The Dissertation Committee for Xue Chen
certifies that this is the approved version of the following dissertation:

Using and Saving Randomness

Committee:

David Zuckerman, Supervisor

Dana Moshkovitz

Eric Price

Yuan Zhou

Using and Saving Randomness

by

Xue Chen

DISSERTATION

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

THE UNIVERSITY OF TEXAS AT AUSTIN

May 2018

Dedicated to my parents Daiguang and Dianhua.

Acknowledgments

First, I am grateful to my advisor, David Zuckerman, for his unwavering support and encouragements. David has been a constant resource for providing me with great advice and insightful feedback about my ideas. Besides these and many fruitful discussions we had, his wisdom, way of thinking, and optimism guided me through the last six years. I also thank him for arranging a visit to the Simons institute in Berkeley, where I enrich my understanding and made new friends. I could not have hoped for a better advisor.

I would like to thank Eric Price. His viewpoints and intuitions opened a different gate for me, which is important for the line of research presented in this thesis. I benefited greatly from our long research meetings and discussions. Besides learning a lot of technical stuff from him, I hope I have absorbed some of his passion and attitudes to research.

Apart from David and Eric, I have had many mentors during my time as a student. I especially thank Yuan Zhou, who has been an amazing collaborator and friend since starting college. I thank Dana Moshkovitz for discussions at various points and being in my thesis committee. I thank Xin Li for many valuable discussions and sharing many research directions. I thank Pinyan Lu for inviting me to China Theory Week and arranging a visit to Shanghai University of Finance and Economics.

I thank the other professors and graduate students in UT Austin. Their discussions and numerous talks give me the chance of learning new results and ideas. I also want to thank my collaborators: Guangda Hu, Daniel Kane, Zhao Song, Xiaoming Sun, and Lei Wang.

Last but the most important of all, I would like to thank my parents for a lifetime of support.

Using and Saving Randomness

Xue Chen, Ph.D.
The University of Texas at Austin, 2018

Supervisor: David Zuckerman

Randomness is ubiquitous and exceedingly useful in computer science. For example, in sparse recovery, randomized algorithms are more efficient and robust than their deterministic counterparts. At the same time, because random sources from the real world are often biased and defective with limited entropy, high-quality randomness is a precious resource. This motivates the studies of pseudorandomness and randomness extraction. In this thesis, we explore the role of randomness in these areas. Our research contributions broadly fall into two categories: learning structured signals and constructing pseudorandom objects.

Learning a structured signal. One common task in audio signal processing is to compress an interval of observation through finding the dominating k frequencies in its Fourier transform. We study the problem of learning a Fourier-sparse signal from noisy samples, where $[0, T]$ is the observation interval and the frequencies can be "off-grid". Previous methods for this problem required the gap between frequencies to be above $1/T$, which is necessary to robustly identify individual frequencies. We show that this gap is not necessary to recover the signal as a whole: for arbitrary k-Fourier-sparse signals under ℓ_2 bounded noise, we provide a learning algorithm with a constant factor growth of the noise and sample complexity polynomial in k and logarithmic in the bandwidth and signal-to-noise ratio.

In addition to this, we introduce a general method to avoid a condition number depending on the signal family $\mathcal{F}$ and the distribution D of measurement in the sample

complexity. In particular, for any linear family $\mathcal{F}$ with dimension d and any distribution D over the domain of $\mathcal{F}$, we show that this method provides a robust learning algorithm with $O(d \log d)$ samples. Furthermore, we improve the sample complexity to $O(d)$ via spectral sparsification (optimal up to a constant factor), which provides the best known result for a range of linear families such as low degree multivariate polynomials. Next, we generalize this result to an active learning setting, where we get a large number of unlabeled points from an unknown distribution and choose a small subset to label. We design a learning algorithm optimizing both the number of unlabeled points and the number of labels.

Pseudorandomness. Next, we study hash families, which have simple forms in theory and efficient implementations in practice. The size of a hash family is crucial for many applications such as derandomization. In this thesis, we study the upper bound on the size of hash families to fulfill their applications in various problems. We first investigate the number of hash functions to constitute a randomness extractor, which is equivalent to the degree of the extractor. We present a general probabilistic method that reduces the degree of any given strong extractor to almost optimal, at least when outputting few bits. For various almost universal hash families including Toeplitz matrices, Linear Congruential Hash, and Multiplicative Universal Hash, this approach significantly improves the upper bound on the degree of strong extractors in these hash families. Then we consider explicit hash families and multiple-choice schemes in the classical problems of placing balls into bins. We construct explicit hash families of almost-polynomial size that derandomizes two classical multiple-choice schemes, which match the maximum loads of a perfectly random hash function.

Table of Contents

List of Tables

List of Figures

List of Algorithms

Chapter 1

Introduction

In this thesis, we study the role of randomness in designing algorithms and constructing pseudorandom objects. Often, randomized algorithms are simpler and faster than their deterministic counterparts. We explore the advantage of randomized algorithms in providing more efficient sample complexities and robust guarantees than deterministic algorithms in learning. On the other hand, both from a theoretic viewpoint and for practical applications, it is desirable to derandomize them and construct pseudorandom objects such as hash functions using as few random bits as possible. We investigate the following two basic questions in computer science:

1. A fundamental question in many fields is to efficiently recover a signal from noisy observations. This problem takes many forms, depending on the measurement model, the signal structure, and the desired norms. For example, coding theory studies the recovery of discrete signals in the Hamming distance. In this thesis, we consider continuous signals that is approximately sparse in the Fourier domain or close to a linear family such as multivariate low degree polynomials. This is a common form in practice. For example, the compression of audio and radio signals exploits their sparsity in the Fourier domain. Often the sample complexity is of more interest than its running rime, because it is costly to take measurements in practical applications. Also, a sample of labeled data requires expensive human/oracle annotation in active learning. This raises a natural question: how can we learn a structured signal from a few noisy

queries? In the first half of this thesis, we design several robust learning algorithms with efficient sample complexity for various families of signals.

2. While randomness is provably necessary for certain tasks in cryptography and distributed computing, the situation is unclear for randomized algorithms. In the second half, we consider an important pseudorandom object in derandomization — hash functions, which have wide applications both in practice and theory. For many problems of derandomization, a key tool is a hash family of small size such that a deterministic algorithm could enumerate all hash functions in this family. In this thesis, we study how to construct small hash families to fulfill their requirements in building randomness extractors and the classical problems of placing balls into bins.

Next we discuss the problems studied in this thesis in detail and show how our methods apply to them. We remark that several techniques developed in this thesis are very general and apply to many other problems.

1.1 Overview

1.1.1 Continuous Sparse Fourier Transforms.

The Fourier transform is a ubiquitous computational tool in processing a variety of signals, including audio, image, and video. In many practical applications, the main reason for using the Fourier transform is that the transformed signal is sparse — it concentrates its energy on k frequencies for a small number k. The sparse representation in the Fourier basis exhibits structures that could be exploited to reduce the sample complexity and speed up the computation of the Fourier transform. For n-point *discrete* signals, this idea has led to a number of efficient algorithms on discrete sparse Fourier transforms, which could achieve the optimal $O(k \log \frac{n}{k})$ sample complexity [IK14] and $O(k \log n \log \frac{n}{k})$ running time [HIKP12].

However, many signals, including audio and radio, are originally from a *continuous* domain. Because the standard way to convert a continuous Fourier transform into a discrete one blows up the sparsity, it is desirable to directly solve the sparse Fourier transform in the *continuous* setting for efficiency gains. In this thesis, we study sparse Fourier transforms in the continuous setting.

Previous work. Let $[0, T]$ be the observation interval of k-Fourier-sparse signals and F denote the "bandlimit", which is the limiting of the frequency domain. This problem, recovering a signal with sparse Fourier transform off the grid (not multiplies of $1/T$), has been a question of extensive study. All previous research starts with finding the frequencies of the signal and then recovers the magnitude of each frequencies.

The first algorithm was by Prony in 1795, which worked in the noiseless setting. Its refinements MUSIC [Sch81] and ESPRIT [RPK86] empirically work better than Prony's algorithm with noise. Matrix pencil [BM86] is a method for computing the maximum likelihood signal under Gaussian noise and evenly spaced samples. Moitra [Moi15] showed that it has an $\mathrm{poly}(k)$ approximation factor if the frequency gap is at least $1/T$.

However, the above results use FT samples, which is analogous to n in the discrete setting. A variety of works [FL12, BCG+14, TBR15] have studied how to adapt sparse Fourier techniques from the discrete setting to get sublinear sample complexity; they all rely on the minimum separation among the frequencies to be at least c/T for $c \geq 1$ or even larger gaps $c = \Omega(k)$ and additional assumptions such as i.i.d. Gaussian noise. Price and Song [PS15] gave the first algorithm with $O(1)$ approximation factor for arbitrary noise, finding the frequencies when $c \gtrsim \log(1/\delta)$, and the signal when $c \gtrsim \log(1/\delta) + \log^2 k$.

All of the above algorithms are designed to recover the frequencies and show this yields a good approximation to the overall signal. Such an approach necessitates $c \geq 1$:

Moitra [Moi15] gave a lower bound, showing that any algorithm finding the frequencies with approximation factor $2^{o(k)}$ must require $c \geq 1$.

Our contribution. In joint work with Kane, Price, and Song [CKPS16], we design the first sample-efficient algorithm that recovers signals with arbitrary k Fourier frequencies from noisy samples. In contrast to all previous approaches that require a frequency gap at least $1/T$ to recover every individual frequency, our algorithm demonstrates that the gap is not necessary to learn the signal: it outputs a sparse representation in the Fourier basis for arbitrary k-Fourier-sparse signals under ℓ_2 bounded noise. This, for the first time, provides a strong theoretic guarantee for the learning of k-Fourier-sparse signals with continuous frequencies.

An important ingredient in our work is the first bound on the condition number $\frac{\|f\|_\infty^2}{\|f\|_2^2} = O(k^4 \log^3 k)$ for any k-Fourier-sparse signal f over the interval $[0, T]$. In this thesis, we present an improved condition number $O(k^3 \log^2 k)$. The condition number of k-sparse signals with $1/T$-separated frequencies is $O(k)$ from the Hilbert inequality. However, for signals with k arbitrary close frequencies, this family is not very well-conditioned — its condition number is $\Omega(k^2)$ from degree $k-1$ polynomials by a Taylor expansion.

On the other hand, degree k polynomials are the special case of k-Fourier-sparse signals in the limit of all frequencies close to 0, by a Taylor expansion. This is a regime with no frequency gap, so previous sparse Fourier results would not apply but our results shows that poly(k) samples suffices. In this special case, we prove that $O(k)$ samples is enough to learn any degree k polynomial under noise, which is optimal up to a constant factor. Our strategy is to query points according to the Chebyshev distribution even though the distance measurement is the uniform distribution over $[0, T]$. A natural question is to generalize this idea to Fourier-sparse signals and avoid the condition number $\Omega(k^2)$. This motivates the follow-up work with Eric Price [CP17].

1.1.2 Condition-number Free Query and Active Learning.

In joint work with Price [CP17], we proposed a framework to avoid the condition number in the sample complexity of learning structured signals. Let $\mathcal{F}$ be a family of signals being learned and D be the distribution to measure the ℓ_2 distance between different signals. We show a strong lower bound on the number of random samples from D to recover a signal in $\mathcal{F}$: the sample complexity depends on the condition number $\sup_{x\in\mathrm{supp}(D)}\sup_{f\in\mathcal{F}}\{\frac{|f(x)|^2}{\mathbb{E}_{y\sim D}[|f(y)|^2]}\}$, which could be arbitrary large under adversarial distributions.

We consider two alternative models of access to the signal to circumvent the lower bound. The first model is that we may *query* the signal where we want, such as in the problem of sparse Fourier transforms. We show how to improve the condition number by biasing the choices of queries towards points of high variance. For a large class of families, this approach significantly saves the number of queries in agnostic learning, including the family of k-Fourier-sparse signals and any linear family.

For continuous sparse Fourier transforms discussed in Section 1.1.1, we present an explicit distribution that scales down the condition number from $O(k^3\log^2 k)$ to $O(k\log^2 k)$ through this approach. Because the condition number is at least k for any query strategy, our estimation $O(k\log^2 k)$ is almost optimal up to log factors. Based on this, we show that $O(k^4\log^3 k + k^2\log^2 k\log FT)$ samples are sufficient to recover any k-Fourier-sparse signal with "bandlimit" F on the interval $[0,T]$, which significantly improves the sample complexity in the previous work [CKPS16].

For any linear family $\mathcal{F}$ of dimension d and any distribution D, we show that this approach scales down the condition number to d and provides an algorithm with $O(d\log d)$ queries to learn *any* signal in $\mathcal{F}$ under arbitrary noise. The $\log d$ factor in the query complexity is due to the fact that the algorithm only uses the distribution with the least condition number to generate all queries. Then, we improve this query complexity to $O(d)$ by designing a sequence of distributions and generating one query from each distribution. On

the other hand, we prove an information lower bound on the number of queries matching the query complexity of our algorithm up to a constant factor. Hence this work completely characterizes query access learning of signals from linear families.

The second model of access is active learning, where the algorithm receives a set of unlabeled points from the *unknown* distribution D and chooses a subset of these points to obtain their labels from noisy observations. We design a learning algorithm that optimizes both the number of labeled and unlabeled examples required to learn a signal from any linear family $\mathcal{F}$ under any unknown distribution D.

Previous work. We notice that the distribution with the least condition number for linear families described in our work [CP17] is similar to strategies proposed in importance sampling [Owe13], leverage score sampling [DMM08, MI10, Woo14], and graph sparsification [SS11]. One significant body of work on sampling regression problems uses leverage score sampling [DMM08, MI10, Mah11, Woo14]. For linear function classes, our sample distribution with the least condition number can be seen as a continuous limit of the leverage score sampling distribution, so our analysis of it is similar to previous work. At least in the fixed design setting, it was previously known that $O(d \log d + d/\varepsilon)$ samples suffice [Mah11].

Multiple researchers [BDMI13, SWZ17] have studied switching from leverage score sampling—which is analogous to the near-linear spectral sparsification of [SS11]—to the linear-size spectral sparsification of [BSS12]. This improves the sample complexity to $O(d/\varepsilon)$, but the algorithms need to know all the y_i as well as the x_i before deciding which x_i to sample; this makes them unsuitable for active/query learning settings. Because [BSS12] is deterministic, this limitation is necessary to ensure the adversary doesn't concentrate the noise on the points that will be sampled. Our result uses the alternative linear-size spectral sparsification of [LS15] to avoid this limitation: we only need the x_i to choose $O(d/\varepsilon)$ points that will perform well (on average) regardless of where the adversary places the noise.

The above references all consider the fixed design setting, while [SM14] gives a method for active regression in the random design setting. Their result shows (roughly speaking) that $O((d \log d)^{5/4} + d/\varepsilon)$ samples of labeled data suffice for the desired guarantee. They do not give a bound on the number of unlabeled examples needed for the algorithm. (Nor do the previous references, but the question does not make sense in fixed-design settings.)

Less directly comparable to this work is [CKNS15], where the noise distribution $(Y \mid x)$ is assumed to be known. This assumption, along with a number of assumptions on the structure of the distributions, allows for significantly stronger results than are possible in our agnostic setting. The optimal sampling distribution is quite different, because it depends on the distribution of the noise as well as the distribution of x.

1.1.3 Existence of Extractors from Simple Hash Families.

Next we consider hash families and their application to construct another pseudorandom object — randomness extractors. In the real world, random sources are often biased and defective, which rarely satisfy the requirements of its applications in algorithm design, distributed computing, and cryptography. A randomness extractor is an efficient algorithm that converts a "weak random source" into an almost uniform distribution. As is standard, we model a weak random source as a probability distribution with min-entropy.

Definition 1.1.1. *The min-entropy of a random variable X is $H_\infty(X) = \min_{x \in \mathrm{supp}(X)} \log_2 \frac{1}{\Pr[X=x]}$.*

It is impossible to construct a deterministic randomness extractor for all sources of min-entropy k [SV86], even if k is as large as $n-1$. Therefore a seeded extractor also takes as input an additional independent uniform random string, called a seed, to guarantee that the output is close to uniform [NZ96].

Definition 1.1.2. *For any $d \in \mathbb{N}^+$, let U_d denote the uniform distribution over $\{0,1\}^d$. For two random variables W and Z with the same support, let $\|W - Z\|$ denote the statistical*

(variation) distance $\|W - Z\| = \max_{T \subseteq \mathrm{supp}(W)} \left| \Pr_{w \sim W}[w \in T] - \Pr_{z \sim Z}[z \in T] \right|$.

$\mathsf{Ext} : \{0,1\}^n \times \{0,1\}^t \to \{0,1\}^m$ *is a* (k, ϵ)*-extractor if for any source* X *with min-entropy* k *and an independent uniform distribution* Y *on* $\{0,1\}^t$, $\|\mathsf{Ext}(X,Y) - U_m\| \leq \epsilon$. *It is a strong* (k, ϵ)*-extractor if in addition, it satisfies* $\left\| \big(\mathsf{Ext}(X,Y), Y\big) - \big(U_m, Y\big) \right\| \leq \epsilon$.

We call 2^t the degree of Ext, because when Ext is viewed as a bipartite graph on $\{0,1\}^n \cup \{0,1\}^m$, its left degree is 2^t. Often the degree 2^t is of more interest than the seed length t. For example, when we view an extractor as a hash family $\mathcal{H} = \big\{\mathsf{Ext}(\cdot, y) | y \in \{0,1\}^t\big\}$, the degree 2^t corresponds to the size of $\mathcal{H}$.

Minimizing the degree of an extractor is crucial for many applications such as constructing optimal samplers and simulating probabilistic algorithms with weak random sources. It is well known that the optimal degree of (k, ϵ)-extractors is $\Theta(\frac{n-k}{\epsilon^2})$, where the upper bound is from the probabilistic method and the lower bound was shown by Radhakrishnan and Ta-Shma [RT00]. At the same time, explicit constructions [Zuc07] with an optimal degree, even for constant error, have a variety of applications in theoretical computer science such as hardness of inapproximability [Zuc07] and constructing almost optimal Ramsey graphs [BDT17].

Most known extractors are sophisticated based on error-correcting codes or expander graphs and complicated to implement in practice. This raises natural questions — are there simple constructions like linear transformations and even simpler ones of Toeplitz matrices? Are there extractors with good parameters and efficient implementations? In joint work with Zuckerman, we answer these questions in the context of the extractor degree. Our main result indicates the existence of strong extractors of linear transformations and Toeplitz matrices and extremely efficient strong extractors with degree close to optimal, at least when outputting few bits.

Our contributions. In joint work with Zuckerman [CZ18], we present a probabilistic construction to improve the degree of any given extractor. We consider the following method

to improve the degree of any strong extractor while keeping almost the same parameters of min entropy and error.

Definition 1.1.3. *Given an extractor* $\mathsf{Ext} : \{0,1\}^n \times \{0,1\}^t \to \{0,1\}^m$ *and a sequence of seeds* $(y_1, \cdots, y_D)$ *where each* $y_i \in \{0,1\}^t$*, we define the restricted extractor* $\mathsf{Ext}_{(y_1,\cdots,y_D)}$ *to be* Ext *restricted in the domain* $\{0,1\}^n \times [D]$ *where* $\mathsf{Ext}_{(y_1,\cdots,y_D)}(x,i) = \mathsf{Ext}(x, y_i)$*.*

Our main result is that given any strong (k,ϵ)-extractor Ext, most restricted extractors with a quasi-linear degree $\tilde{O}(\frac{n}{\epsilon^2})$ from Ext are strong $(k, 3\epsilon)$-extractors for a constant number of output bits, despite the degree of Ext.

Theorem 1.1.4. *There exists a universal constant* C *such that given any strong* (k,ϵ)*-extractor* $\mathsf{Ext} : \{0,1\}^n \times \{0,1\}^t \to \{0,1\}^m$*, for* $D = C \cdot \frac{n \cdot 2^m}{\epsilon^2} \cdot \log^2 \frac{n \cdot 2^m}{\epsilon}$ *random seeds* $y_1, \ldots, y_D \in \{0,1\}^t$*,* $\mathsf{Ext}_{(y_1,\ldots,y_D)}$ *is a strong* $(k, 3\epsilon)$*-extractor with probability 0.99.*

We state the corollaries of Theorem 1.1.4 for simple extractors of linear transformations and Toeplitz matrices and hash families with efficient implementations in Chapter 7.

On the other hand, the same statement of Theorem 1.1.4 holds for extractors. In this case, we observe that the dependency 2^m on the degree D of restricted extractors is necessary to guarantee its error is less than $1/2$ on m output bits.

Proposition 1.1.5. *There exists a* (k,ϵ)*-extractor* $\mathsf{Ext} : \{0,1\}^n \times \{0,1\}^t \to \{0,1\}^m$ *with* $k = 1$ *and* $\epsilon = 0$ *such that any restricted extractor of* Ext *requires the degree* $D \geq 2^{m-1}$ *to guarantee its error is less than* $1/2$*.*

While the dependency 2^m is necessary for the degree of restricted extractors, it may not be necessary for *strong* extractors.

Previous work. In a seminal work, Impagliazzo, Levin, and Luby [ILL89] proved the Leftover Hash Lemma, i.e., all functions from an almost universal hash family constitute a strong extractor. In particular, this implies that all linear transformations and all Toeplitz matrices constitute strong extractors respectively.

Most previous research has focused on linear extractors, whose extractor functions are linear on the random source for every fixed seed. Because of their simplicity and various applications such as building blocks of extractors for structured sources [Li16, CL16], there have been several constructions of linear extractors with small degree. The first nontrivial progress was due to Trevisan [Tre01], who constructed the first linear extractor with degree polynomial in n and ϵ. Based on Trevisan's work, Shaltiel and Umans [SU05] built linear extractors with almost linear degree for constant error. Later on, Guruswami, Umans, and Vadhan [GUV09a] constructed almost optimal linear condensers and vertex-expansion expanders, which are variants of extractors and lead to an extractor with a degree $n \cdot \mathrm{poly}(k/\epsilon)$. However, the GUV extractor is not linear. Moreover, it is still open whether the degree of linear extractors could match the degree $\Theta(\frac{n-k}{\epsilon^2})$ of general extractors.

On the other hand, much less is known about extractors consisting of Toeplitz matrices. Prior to this work, even for $m = 2$ output bits, the best known upper bound on extractors with Toeplitz matrices was exponential in n by the Leftover Hash Lemma [ILL89] of all Toeplitz matrices.

At the same time, from a practical point of view, it is desirable to have an extractor with a small degree that runs fast and is easy to implement. In this work, we consider efficient extractors from almost universal hash families, which are easier to implement than error-correcting codes and expander graphs used in most known constructions of extractors. In Chapter 7, we describe a few notable almost universal hash families with efficient implementations. Prior to this work, the best known upper bounds on the degree of extractors from almost universal hash families were the Leftover Hash Lemma [ILL89], which are

exponential in the length of the random source n.

Other work on improving extractors focus on other parameters such as the error and number of output bits. Raz et al. [RRV99] showed how to reduce the error and enlarge the number of output bits of any given extractor by sacrificing the degree.

1.1.4 Hash functions for Multiple-choice Schemes.

We consider explicit hash families for the classical problem of placing balls into bins. The basic model is to hash n balls into n bins independently and uniformly at random, which we call the 1-choice scheme. A well-known and useful fact of the 1-choice scheme is that with high probability, each bin contains at most $O(\frac{\log n}{\log \log n})$ balls. By high probability, we mean probability $1 - n^{-c}$ for an arbitrary constant c.

An alternative variant, which we call *Uniform-Greedy*, is to provide $d \geq 2$ independent random choices for each ball and place the ball in the bin with the lowest load. In a seminal work, Azar et al. [ABKU99] showed that the *Uniform-Greedy* scheme with d independent random choices guarantees a maximum load of only $\frac{\log \log n}{\log d} + O(1)$ with high probability for n balls. Later, Vöcking [Voc03] introduced the *Always-Go-Left* scheme to further improve the maximum load to $\frac{\log \log n}{d \log \phi_d} + O(1)$ for d choices where $\phi_d > 1.61$ is the constant satisfying $\phi_d^d = 1 + \phi_d + \cdots + \phi_d^{d-1}$. For convenience, we always use d-choice schemes to denote the *Uniform-Greedy* and *Always-Go-Left* scheme with $d \geq 2$ choices.

Traditional analysis of load balancing assumes a perfectly random hash function. A large body of research is dedicated to the removal of this assumption by designing explicit hash families using fewer random bits. In the 1-choice scheme, it is well known that $O(\frac{\log n}{\log \log n})$-wise independent functions guarantee a maximum load of $O(\frac{\log n}{\log \log n})$ with high probability, which reduces the number of random bits to $O(\frac{\log^2 n}{\log \log n})$. Celis et al. [CRSW13] designed a hash family with a description of $O(\log n \log \log n)$ random bits that achieves the same maximum load of $O(\frac{\log n}{\log \log n})$ as a perfectly random hash function.

In this thesis, we are interested in the explicit constructions of hash families that achieve *the same maximum loads* as a perfectly random hash function in the d-choice schemes. More precisely, we study how to derandomize the perfectly random hash function in the *Uniform-Greedy* and *Always-Go-Left* scheme. For these two schemes, $O(\log n)$-wise independent hash functions achieve the same maximum loads from Vöcking's argument [Voc03], which provides a hash family with $\Theta(\log^2 n)$ random bits. Recently, Reingold et al. [RRW14] showed that the hash family in [CRSW13] guarantees a maximum load of $O(\log \log n)$ in the *Uniform-Greedy* scheme with $O(\log n \log \log n)$ random bits.

Our results. We construct a hash family with $O(\log n \log \log n)$ random bits based on the previous work of Celis et al. [CRSW13] and show the following results.

1. This hash family has a maximum load of $\frac{\log \log n}{\log d} + O(1)$ in the *Uniform-Greedy* scheme.
2. It has a maximum load of $\frac{\log \log n}{d \log \phi_d} + O(1)$ in the *Always-Go-Left* scheme.

The maximum loads of our hash family match the maximum loads of a perfectly random hash function [ABKU99, Voc03] in the *Uniform-Greedy* and *Always-Go-Left* scheme separately.

1.1.5 CSPs with a global cardinality constraint.

A variety of problems in pseudorandomness such as bipartite expanders and dispersers could be stated as constraint satisfaction problems complying with a global cardinality constraint. In a d-ary constraint satisfaction problem (CSP), we are given a set of boolean variables $\{x_1, x_2, \cdots, x_n\}$ over $\{\pm 1\}$ and m constraints $C_1, \cdots, C_m$, where each constraint C_i consists of a predicate on at most d variables. A constraint is satisfied if and only if the assignment of the related variables is in the predicate of the constraint. The task is to find an assignment to $\{x_1, \cdots, x_n\}$ so that the greatest (or the least) number of constraints in $\{C_1, \cdots, C_m\}$ are satisfied.

Given a boolean CSP instance $\mathcal{I}$, we can impose a global cardinality constraint $\sum_{i=1}^{n} x_i = (1-2p)n$ (we assume that pn is an integer). Such a constraint is called the *bisection constraint* if $p = 1/2$. For example, the MaxBisection problem is the MaxCut problem with the bisection constraint. Constraint satisfaction problems with global cardinality constraints are natural generalizations of boolean CSPs. Researchers have been studying approximation algorithms for CSPs with global cardinality constraints for decades, where the MaxBisection problem [FJ95, Ye01, HZ02, FL06, GMR+11, RT12, ABG13] and the SmallSet Expansion problem [RS10, RST10, RST12] are two prominent examples.

Adding a global cardinality constraint could strictly enhance the hardness of the problem. The SmallSet Expansion problem can be viewed as the MinCUT problem with the cardinality of the selected subset to be $\rho|V|$ ($\rho \in (0,1)$). While MinCUT admits a polynomial-time algorithm to find the optimal solution, we do not know a good approximation algorithm for SmallSet Expansion. Raghavendra and Steurer [RS10] suggested that the SmallSet Expansion problem is where the hardness of the notorious UniqueGames problem [Kho02] stems from.

We study several problems about CSPs under a global cardinality constraint in this thesis. The first one is the fixed parameter tractability (FPT) of the CSP above average under a global cardinality constraint. Specifically, let AVG be the expected number of constraints satisfied by randomly choosing an assignment to $x_1, x_2, \ldots, x_n$ complying with the global cardinality constraint. We show an efficient algorithm that finds an assignment (complying with the cardinality constraint) satisfying more than $(AVG+t)$ constraints for an input parameter t. The second is to approximate the vertex expansion of a bipartite graph. Our main results are strong integrality gaps in the Lasserre hierarchy and an approximation algorithm for dispersers and bipartite expanders.

1.2 Organization

We provide some preliminaries and discuss condition numbers in Chapter 2. We also introduce a few tools in signal recovery of continuous sparse Fourier transform and linear families in this chapter.

In Chapter 3, we prove the condition number of k-Fourier-sparse signals is in $[k^2, \tilde{O}(k^3)]$ and show how to scale it down to $\tilde{O}(k)$. In Chapter 4, we present the sample-efficient algorithm for continuous sparse Fourier transform. In Chapter 5, we consider a special case of continuous sparse Fourier transform and present the robust polynomial recovery algorithm. In Chapter 6, we study how to learn signals from any given linear families. Most of the material in these chapters are based on joint work with Kane, Price, and Song [CKPS16, CP17].

We consider hash functions and its applications in Chapters 7 and 8. We show how to reduce the degree of any extractor and discuss its implications on almost universal hash families in Chapter 7. We present our hash families that derandomizes multiple-choice schemes in Chapter 8. Most of the material in these chapters are based on joint work with Zuckerman [CZ18] and [Che17].

We study problems about CSPs under a global cardinality constraint in Chapters 9 and 10. We show that they are fixed parameter tractable in Chapter 9. We present the integrality gaps of approximating dispersers and bipartite expanders in Chapter 10. Most of the material in these chapters are based on joint work with Zhou [CZ17] and [Che16].

Chapter 2

Preliminaries

Let $[n] = \{1, 2, \cdots, n\}$. For convenience, we always use $\binom{S}{d}$ to denote the set of all subsets of size d in S and $\binom{S}{\leq d}$ to denote the set of all subsets of size at most d in S (including $\emptyset$). For two subsets S and T, we use $S \Delta T$ to denote the symmetric difference of S and T.

Let $n!$ denote the product $\prod_{i=1}^{n} i$ and $n!! = \prod_{i=0}^{[\frac{n-1}{2}]}(n-2i)$. We use $\vec{0}$ ($\vec{1}$ resp.) to denote the all 0 (1 resp.) vector and 1_E to denote the indicator variable of an event E, i.e. $1_E = 1$ when E is true, and $1_E = 0$ otherwise.

We use $X \lesssim Y$ to denote the inequality $X \leq C \cdot Y$ for a universal constant C. For a random variable X on $\{0,1\}^n$ and a function f from $\{0,1\}^n$, let $f(X)$ denote the random variable $f(x)$ on the image of f when $x \sim X$.

Given a distribution D, let $\|f\|_D$ denote the ℓ_2 norm in D, i.e., $(\mathop{\mathbb{E}}_{x \sim D}[|f(x)|^2])^{1/2}$. Given a sequence $S = (t_1, \ldots, t_m)$ (allowing repetition in S) and corresponding weights $(w_1, \ldots, w_m)$, let $\|f\|_{S,w}^2$ denote the weighted ℓ_2 norm $\sum_{j=1}^{m} w_j \cdot |f(t_j)|^2$. For convenience, we omit w if it is a uniform distribution on S, i.e., $\|f\|_S = \left(\mathbb{E}_{i \in [m]}\left[|f(t_i)|^2\right]\right)^{1/2}$. For a vector $\vec{v} = \big(v(1), \ldots, v(m)\big) \in \mathbb{C}^m$, let $\|\vec{v}\|_k$ denote the L_k norm, i.e., $\left(\sum_{i \in [m]} |v(i)|^k\right)^{1/k}$.

Given a matrix $A \in \mathbb{C}^{m \times m}$, let A^* denote the conjugate transpose $A^*(i,j) = \overline{A(j,i)}$ and $\|A\|$ denote the operator norm $\|A\| = \max_{\vec{v} \neq \vec{0}} \frac{\|A\vec{v}\|_2}{\|\vec{v}\|_2}$ and $\lambda(A)$ denote all eigenvalues of A. Given a self-adjoint matrix A, let $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ denote the smallest eigenvalue and the largest eigenvalue of A.

Then we discuss the condition numbers of a family under D.

2.1 Condition Numbers

Given a family $\mathcal{F}$ (not necessarily linear) of signals and a fixed distribution D on the domain of $\mathcal{F}$, we consider the problem of estimating $\|f\|_D^2$ with high confidence and define the worst-case condition number and average condition number of $\mathcal{F}$ under D.

Let $x_1, \cdots, x_m$ be m random samples from D. The basic way to estimate $\|f\|_D^2 = \mathop{\mathbb{E}}_{x\sim D}\left[|f(x)|^2\right]$ is $\frac{1}{m}\sum_{i=1}^m |f(x_i)|^2$. To show that this concentrates, we would like to apply Chernoff bounds, which depend on the maximum value of summand. In particular, the concentration depends on the *worst-case condition number* of $\mathcal{F}$, i.e.,

$$K_{\mathcal{F}} = \sup_{x\in\mathrm{supp}(D)} \sup_{f\in\mathcal{F}} \frac{|f(x)|^2}{\|f\|_D^2}.$$

We consider estimating $\|f\|_D^2 = \mathop{\mathbb{E}}_{x\sim D}\left[|f(x)|^2\right]$ by samples from any other distribution D'. For any distribution D' over $\mathrm{supp}(D)$, for m samples $t_1, \cdots, t_m$ from D', we always assign the weight $w_i = \frac{D(x)}{D'(x)\cdot m}$ for each $i \in [m]$ such that

$$\mathop{\mathbb{E}}_{t_1,\cdots,t_m}\left[\sum_{i=1}^m w_i \cdot |f(t_i)|^2\right] = \sum_{i=1}^m \mathop{\mathbb{E}}_{t_i\sim D'}\left[\frac{D(t_i)}{D'(t_i)\cdot m}\cdot |f(t_i)|^2\right] = \mathop{\mathbb{E}}_{t\sim D}\left[|f(t)|^2\right] = \|f\|_D^2.$$

When D is clear, we also use $f^{(D')}(x)$ to denote the weighted function $\sqrt{\frac{D(x)}{D'(x)}}\cdot f(x)$ such that $\mathop{\mathbb{E}}_{x\sim D}\left[|f(x)|^2\right] = \mathop{\mathbb{E}}_{x\sim D'}\left[|f^{(D')}(x)|^2\right]$. When $\mathcal{F}$ and D are clear, let $K_{D'}$ denote the condition number of signals in $\mathcal{F}$ from random sampling of D':

$$K_{D'} = \sup_{x\in\mathrm{supp}(D)} \sup_{f\in\mathcal{F}} \frac{D(x)}{D'(x)}\cdot\frac{|f(x)|^2}{\|f\|_D^2} = \sup_{x\in\mathrm{supp}(D)}\left\{\frac{D(x)}{D'(x)}\cdot\sup_{f\in\mathcal{F}}\frac{|f(x)|^2}{\|f\|_D^2}\right\}.$$

Let $D_{\mathcal{F}}$ be the distribution minimizing $K_{D'}$ by making the inner term the same for every x, i.e.,

$$D_{\mathcal{F}}(x) = \frac{D(x)\cdot\sup_{f\in\mathcal{F}}\frac{|f(x)|^2}{\|f\|_D^2}}{\kappa_{\mathcal{F}}} \text{ for } \kappa_{\mathcal{F}} = \mathop{\mathbb{E}}_{x\sim D}\left[\sup_{f\in\mathcal{F}}\frac{|f(x)|^2}{\|f\|_D^2}\right]. \tag{2.1}$$

For convenience, we call $\kappa_{\mathcal{F}}$ the *average condition number* of $\mathcal{F}$. It is straightforward to verify the condition number $K_{D_{\mathcal{F}}} = \kappa_{\mathcal{F}}$ for $\mathcal{F}$ with random samples from $D_{\mathcal{F}}$.

Claim 2.1.1. *For any family $\mathcal{F}$ and any distribution D on its domain, let $D_{\mathcal{F}}$ be the distribution defined in (2.1) with $\kappa_{\mathcal{F}}$. The condition number $K_{D_{\mathcal{F}}} = \sup_x \left\{ \sup_{f \in \mathcal{F}} \{ \frac{D(x)}{D_{\mathcal{F}}(x)} \cdot \frac{|f(x)|^2}{\|f\|_D^2} \} \right\}$ is at most $\kappa_{\mathcal{F}}$.*

Proof. For any $g \in \mathcal{F}$ and x in the domain G,

$$\frac{|g(x)|^2}{\|g\|_D^2} \cdot \frac{D(x)}{D_{\mathcal{F}}(x)} = \frac{\frac{|g(x)|^2}{\|g\|_D^2} \cdot D(x)}{\sup_{f \in \mathcal{F}} \frac{|f(x)|^2}{\|f\|_D^2} \cdot D(x)/\kappa_{\mathcal{F}}} \leq \kappa_{\mathcal{F}}.$$

□

In this rest of this work, we will use $\kappa_{\mathcal{F}}$ to denote the condition number of $D_{\mathcal{F}}$. We discuss the application of this approach to sparse Fourier transform and linear families in Chapter 3 and Chapter 6 separately.

2.2 Chernoff Bounds

We state a few versions of the Chernoff bounds for random sampling. We start with the Chernoff bound for real numbers [Che52].

Lemma 2.2.1. *Let $X_1, X_2, \cdots, X_n$ be independent random variables. Assume that $0 \leq X_i \leq 1$ always, for each $i \in [n]$. Let $X = X_1 + X_2 + \cdots + X_n$ and $\mu = \mathbb{E}[X] = \sum_{i=1}^{n} \mathbb{E}[X_i]$. Then for any $\varepsilon > 0$,*

$$\mathsf{Pr}[X \geq (1+\varepsilon)\mu] \leq \exp(-\frac{\varepsilon^2}{2+\varepsilon}\mu) \text{ and } \mathsf{Pr}[X \geq (1-\varepsilon)\mu] \leq \exp(-\frac{\varepsilon^2}{2}\mu).$$

In this work, we use the following version of the Chernoff bound.

Corollary 2.2.2. *Let $X_1, X_2, \cdots, X_n$ be independent random variables in $[0, R]$ with expectation 1. For any $\varepsilon < 1/2$, $X = \frac{\sum_{i=1}^{n} X_i}{n}$ with expectation 1 satisfies*

$$\Pr[|X - 1| \geq \varepsilon] \leq 2\exp(-\frac{\varepsilon^2}{3} \cdot \frac{n}{R}).$$

We state the matrix Chernoff inequality from [Tro12].

Theorem 2.2.3 (Theorem 1.1 of [Tro12]). *Consider a finite sequence $\{X_k\}$ of independent, random, self-adjoint matrices of dimension d. Assume that each random matrix satisfies*

$$X_k \succeq 0 \quad \text{and} \quad \lambda(X_k) \leq R.$$

Define $\mu_{\min} = \lambda_{\min}(\sum_k \mathbb{E}[X_k])$ and $\mu_{\max} = \lambda_{\max}(\sum_k \mathbb{E}[X_k])$. Then

$$\Pr\left\{\lambda_{\min}(\sum_k X_k) \leq (1-\delta)\mu_{\min}\right\} \leq d\left(\frac{e^{-\delta}}{(1-\delta)^{1-\delta}}\right)^{\mu_{\min}/R} \quad \text{for } \delta \in [0,1], \text{and} \tag{2.2}$$

$$\Pr\left\{\lambda_{\max}(\sum_k X_k) \geq (1+\delta)\mu_{\max}\right\} \leq d\left(\frac{e^{-\delta}}{(1+\delta)^{1+\delta}}\right)^{\mu_{\max}/R} \quad \text{for } \delta \geq 0 \tag{2.3}$$

Chapter 3

Condition Numbers of Continuous Sparse Fourier Transform

We study the worst-case and average condition number of k-Fourier-sparse signals in the continuous setting. Without loss of generality, we set $[-1, 1]$ to be the interval of observations and F to be the bandlimit of frequencies. In this chapter, we set the family of k-Fourier-sparse signals as

$$\mathcal{F} = \left\{ f(x) = \sum_{j=1}^{k} v_j \cdot e^{2\pi \boldsymbol{i} f_j x} \middle| v_j \in \mathbb{C}, |f_j| \leq F \right\}. \tag{3.1}$$

We consider the uniform distribution over $[-1, 1]$ and define $\|f\|_2 = \left(\mathop{\mathbb{E}}_{x \in [-1,1]} [|f(x)|^2] \right)^{1/2}$ to denote the energy of a signal f.

We prove lower and upper bounds on the worst-case condition number and average condition number of $\mathcal{F}$. Our main constribution are two upper bounds on the condition numbers of $\mathcal{F}$, which are the first result about the condition numbers of sparse Fourier transform with continuous frequencies.

Theorem 3.0.1. *Let $\mathcal{F}$ be the family defined in (3.1) given F and k.*

$$K_{\mathcal{F}} = \sup_{f \in \mathcal{F}} \sup_{x \in [-1,1]} \frac{|f(x)|^2}{\|f\|_2^2} = O(k^3 \log^3 k).$$

At the same time, there exists $f \in \mathcal{F}$ such that $\sup_{x \in [-1,1]} \frac{|f(x)|^2}{\|f\|_2^2} = k^2$.

Furthermore, we show the average condition number $\kappa_{\mathcal{F}}$ is $\tilde{O}(k)$.

Theorem 3.0.2. *Given any F and k, let $\mathcal{F}$ be the family defined in (3.1).*

$$\kappa_{\mathcal{F}} = \mathop{\mathbb{E}}_{x\in[-1,1]}\left[\sup_{f\in\mathcal{F}}\frac{|f(x)|^2}{\|f\|_2^2}\right] = O(k\log^2 k).$$

From Claim 2.1.1 in Chapter 2, we obtain the following corollary. Recall that $D(x) = 1/2$ for the uniform distribution over $[-1,1]$.

Corollary 3.0.3. *For any k, there exists an explicit distribution $D_{\mathcal{F}}$ such that*

$$\forall x\in[-1,1], \sup_{f\in\mathcal{F}}\left\{\frac{1}{2\cdot D_{\mathcal{F}}(x)}\cdot\frac{|f(x)|^2}{\|f\|_2^2}\right\} = O(k\log^2 k).$$

On the other hand, the average condition number $\kappa_{\mathcal{F}} = \mathop{\mathbb{E}}_{x\in[-1,1]}\left[\sup_{f\in\mathcal{F}}\frac{|f(x)|^2}{\|f\|_2^2}\right]$ is at least k, because for any $x\in[-1,1]$, there exists a periodic signal f with $\frac{|f(x)|^2}{\|f\|_2^2} = k$. This indicates our estimation of average condition number is almost tight up to log factors.

In the rest of this section, we prove the upper bound of Theorem 3.0.1 in Section 3.1. Then we prove Theorem 3.0.2 and Corollary 3.0.3 in Section 3.2. For completeness, we show the lower bound of Theorem 3.0.1 through the approximation of degree $k-1$ polynomials by k-Fourier sparse signals in Section 3.3.

3.1 The Worst-case Condition number of Fourier Sparse Signals

We bound the worst-case condition number of signals in $\mathcal{F}$ in this section. We first state the technical result to prove the upper bound in Theorem 3.0.1.

Theorem 3.1.1. *Given any $k > 0$, there exists $d = O(k^2\log k)$ such that for any $f(x) = \sum_{j=1}^{k} v_j\cdot e^{2\pi \mathbf{i} f_j\cdot x}$, any $t\in\mathbb{R}$, and any $\Delta > 0$,*

$$|f(t)|^2 \le O(k)\cdot(\sum_{j=1}^{d}|f(t+j\cdot\Delta|^2)$$

We finish the proof of Theorem 3.0.1 bounding the worst-case condition number of $K_{\mathcal{F}}$ by the above relation.

Proof of Theorem 3.0.1. Given any $f \in \mathcal{F}$, we prove that

$$|f(t)|^2 = O(k^3 \log^2 k) \int_t^1 |f(x)|^2 \mathrm{d}x \text{ for any } t \le 0,$$

which indicates $|f(t)|^2 = O(k^3 \log^2 k) \cdot \mathop{\mathbb{E}}_{x\sim[-1,1]}\big[|f(x)|^2\big]$. By symmetry, it also implies that $|f(t)|^2 = O(k^3 \log^2 k) \cdot \mathop{\mathbb{E}}_{x\sim[-1,1]}\big[|f(x)|^2\big]$ for any $t \ge 0$.

We use Theorem 3.1.1 on $f(t)$:

$$\begin{aligned}
\frac{1-t}{d} \cdot |f(t)|^2 &\le O(k) \cdot \int_{\Delta=0}^{\frac{1-t}{d}} \sum_{j\in[d]} |f(t+j\Delta)|^2 \mathrm{d}\Delta \\
&\lesssim k \sum_{j\in[d]} \int_{\Delta=0}^{\frac{1-t}{d}} |f(t+j\Delta)|^2 \mathrm{d}\Delta \\
&\lesssim k \sum_{j\in[d]} \frac{1}{j} \cdot \int_{\Delta'=0}^{\frac{(1-t)j}{d}} |f(t+\Delta')|^2 \mathrm{d}\Delta' \\
&\lesssim k \sum_{j\in[d]} \frac{1}{j} \cdot \int_{x=-1}^{1} |f(x)|^2 \mathrm{d}x \\
&\lesssim k \log k \cdot \int_{x=-1}^{1} |f(x)|^2 \mathrm{d}x.
\end{aligned}$$

From all discussion above, we have $|f(t)|^2 \lesssim dk \log k \cdot \mathop{\mathbb{E}}_{x\in[-1,1]}[|f(x)|^2]$. □

We prove Theorem 3.1.1 in the rest of this section. We first provide a polynomial interpolation lemma.

Theorem 3.1.2. *Given $z_1, \cdots, z_k$ with $|z_1| = |z_2| = \cdots = z_k = 1$, there exists a degree $d = O(k^2 \log k)$ polynomial $P(z) = \sum_{j=0}^{d} c(j) \cdot z^j$ satisfying*

1. $P(z_i) = 0$ *for each* $i \in [k]$.

2. *Its coefficients satisfy* $|c(0)|^2 = O(k) \cdot \left(\sum_{j=1}^{d} |c(j)|^2\right)$.

We use residual polynomials to prove Theorem 3.1.2.

Lemma 3.1.3. *Given* $z_1, \cdots, z_k$, *for any integer* n, *let* $r_{n,k}(z) = \sum_{i=0}^{k-1} r_{n,k}^{(i)} \cdot z^i$ *denote the residual polynomial of* $r_{n,k} \equiv z^n \mod \prod_{j=1}^{k}(z - z_j)$. *Then each coefficient in* $r_{n,k}$ *is bounded:* $|r_{n,k}^{(i)}| \leq \binom{k-1}{i} \cdot \binom{n}{k-1}$ *for* $n \geq k$ *and* $|r_{n,k}^{(i)}| \leq \binom{k-1}{i} \cdot \binom{|n|+k-1}{k-1}$ *for* $n < 0$.

For completeness, we provide a proof of Lemma 3.1.3 in Appendix A. We finish the proof of Theorem 3.1.2 here.

Proof. Let C_0 be a large constant and $d = 5 \cdot k^2 \log k$. We use $\mathcal{P}$ to denote the following subset of polynomials with bounded coefficients:

$$\left\{ \sum_{j=0}^{d} \alpha_j \cdot 2^{-j/k} \cdot z^j \middle| \alpha_0, \ldots, \alpha_d \in [-C_0, C_0] \cap \mathbb{Z} \right\}.$$

For each polynomial $P(z) = \sum_{j=0}^{d} \alpha_j \cdot 2^{-j/k} \cdot z^j \in \mathcal{P}$, we rewrite $P(z) \mod \prod_{j=1}^{k}(z - z_j)$ as

$$\sum_{j=0}^{d} \alpha_j \cdot 2^{-j/k} \cdot \left(z^j \mod \prod_{j=1}^{k}(z - z_j) \right) = \sum_{i=0}^{k-1} \left(\sum_{j=0}^{d} \alpha_j \cdot 2^{-j/k} \cdot r_{n,k}^{(i)} \right) z^i.$$

The coefficient $\sum_{j=0}^{d} \alpha_j \cdot 2^{-j/k} \cdot r_{n,k}^{(i)}$ is bounded by

$$\sum_{j=0}^{d} C_0 \cdot 2^{-j/k} \cdot 2^k j^{k-1} \leq d \cdot C_0 \cdot 2^k \cdot d^k \leq d^{2k}.$$

We apply the pigeon hole theorem on the $(2C_0+1)^d$ polynomials in $\mathcal{P}$ after module $\prod_{j=1}^{d}(z - z_j)$: there exists $m > (2C_0 + 1)^{0.9d}$ polynomials $P_1, \cdots, P_m$ such that each coefficient of $(P_i - P_j) \mod \prod_{j=1}^{k}(z - z_j)$ is d^{-2k} small from the counting

$$\frac{(2C_0 + 1)^d}{(d^{2k}/d^{-2k})^{2k}} > (2C_0 + 1)^{0.9d}.$$

Because $m > (2C_0+1)^{0.9d}$, there exists $j_1 \in [m]$ and $j_2 \in [m] \setminus \{j_1\}$ such that the lowest monomial z^l with different coefficients in P_{j_1} and P_{j_2} satisfies $l \le 0.1d$. Eventually we set

$$P(z) = z^{-l} \cdot \big(P_{j_1}(z) - P_{j_2}(z)\big) - \left(z^{-l} \mod \prod_{j=1}^{k}(z-z_j)\right) \cdot \left(P_{j_1}(z) - P_{j_2}(z) \mod \prod_{j=1}^{k}(z-z_j)\right)$$

to satisfy the first property $P(z_1) = P(z_2) = \cdots = P(z_k) = 0$. We prove the second property in the rest of this proof.

We bound every coefficient in $\big(z^{-l} \mod \prod_{j=1}^{k}(z-z_j)\big) \cdot \big(P_{j_1}(z) - P_{j_2}(z) \mod \prod_{j=1}^{k}(z-z_j)\big)$ by $k \cdot 2^l (l+k)^{k-1} \cdot d^{-2k} \le d \cdot 2^d d^{k-1} \cdot d^{-2k} \le d^{-0.5k}$. On the other hand, the constant coefficient in $z^{-l} \cdot \big(P_{j_1}(z) - P_{j_2}(z)\big)$ is at least $2^{-l/k} \ge 2^{-0.1d/k} = k^{-0.5k}$ because z^l is the smallest monomial with different coefficients in P_{j_1} and P_{j_2} from $\mathcal{P}$. Thus the constant coefficient $|C(0)|^2$ of $P(z)$ is at least $0.5 \cdot 2^{-2l/k}$.

Next we upper bound the sum of the rest coefficients $\sum_{j=1}^{d} |C(j)|^2$ by

$$\sum_{j=1}^{d} (2C_0 \cdot 2^{-(l+j)/k} + d^{-0.5k})^2 \le 2 \cdot 4C_0^2 \sum_{j=1}^{d} 2^{-2(l+j)/k} + 2 \cdot \sum_{j=1}^{d} d^{-0.5k \cdot 2} \lesssim k \cdot 2^{-2l/k},$$

which demonstrates the second property. □

Then we finish the proof of Theorem 3.1.1 using the above polynomial interpolation bound.

Proof of Theorem 3.1.1. Given k frequencies $f_1, \cdots, f_k$ and Δ, we set $z_1 = e^{2\pi \boldsymbol{i} f_1 \cdot \Delta}, \cdots, z_k = e^{2\pi \boldsymbol{i} f_k \cdot \Delta}$. Let $C(0), \cdots, C(d)$ be the coefficients of the degree d polynomial $P(z)$ in Theorem 3.1.2. We have

$$\begin{aligned}
\sum_{j=0}^{d} C(j) \cdot f(t + j \cdot \Delta) &= \sum_{j=0}^{d} C(j) \sum_{j' \in [k]} v_{j'} \cdot e^{2\pi \boldsymbol{i} f_{j'}(t + j\Delta)} \\
&= \sum_{j=0}^{d} C(j) \sum_{j' \in [k]} v_{j'} \cdot e^{2\pi \boldsymbol{i} f_{j'} t} \cdot z_{j'}^{j} = \sum_{j' \in [k]} v_{j'} \cdot e^{2\pi \boldsymbol{i} f_{j'} t} \sum_{j=0}^{d} C(j) \cdot z_{j'}^{j} = 0.
\end{aligned}$$

Hence for every $i \in [k]$,

$$-C(0) \cdot f(t) = \sum_{j=1}^{d} C(j) \cdot f(t + j \cdot \Delta). \tag{3.2}$$

By Cauchy-Schwartz inequality, we have

$$|C(0)|^2 \cdot |f(t)|^2 \leq \left(\sum_{j=1}^{d} |C(j)|^2 \right) \cdot \left(\sum_{j=1}^{d} |f(t + j \cdot \Delta)|^2 \right). \tag{3.3}$$

From the second property of $C(0), \cdots, C(d)$ in Theorem 3.1.2, $|f(t)|^2 \leq O(k) \cdot (\sum_{j=1}^{d} |f(t + j \cdot \Delta|^2)$. □

3.2 The Average Condition number of Fourier Sparse Signals

We bound the average condition number of $\mathcal{F}$ in this section. The key ingrediant is a upper bound on $|f(t)|^2/\|f\|_2^2$ *depending on* t.

Lemma 3.2.1. *For any* $t \in (-1, 1)$,

$$\sup_{f \in \mathcal{F}} \frac{|f(t)|^2}{\|f\|_2^2} \lesssim \frac{k \log k}{1 - |t|}.$$

We state the improvement compared to Theorem 3.1.1 bounding the worst-case condition number.

Claim 3.2.2. *Given* $f(x) = \sum_{j=1}^{k} v_j e^{2\pi \boldsymbol{i} f_j \cdot x}$ *and* Δ, *there exists* $l \in [2k]$ *such that for any* t,

$$|f(t + l \cdot \Delta)|^2 \lesssim \sum_{j \in [2k] \setminus \{l\}} |f(t + j \cdot \Delta)|^2.$$

Proof. Given k frequencies $f_1, \cdots, f_k$ and Δ, we set $z_1 = e^{2\pi \boldsymbol{i} f_1 \cdot \Delta}, \cdots, z_k = e^{2\pi \boldsymbol{i} f_k \cdot \Delta}$. Let V be the linear subspace

$$\left\{ \big(\alpha(0), \ldots, \alpha(2k-1)\big) \in \mathbb{C}^{2k} \Big| \sum_{j=0}^{2k-1} \alpha(j) \cdot z_i^j = 0, \forall i \in [k] \right\}.$$

Because the dimension of V is k, let $\alpha_1, \cdots, \alpha_k \in V$ be k orthogonal coefficient vectors with unit length $\|\alpha_i\|_2 = 1$. Let l be the coordinate in $[2k]$ with the largest weight $\sum_{i=1}^{k} |\alpha_i(l)|^2$. We prove the main technical result.

From the definition of α_i, we have

$$\begin{aligned}\sum_{j\in[2k]} \alpha_i(j) \cdot f(t + j \cdot \Delta) &= \sum_{j\in[2k]} \alpha_i(j) \sum_{j'\in[k]} v_{j'} \cdot e^{2\pi \boldsymbol{i} f_{j'} \cdot (t+j\Delta)} \\ &= \sum_{j\in[2k]} \alpha_i(j) \sum_{j'\in[k]} v_{j'} \cdot e^{2\pi \boldsymbol{i} f_{j'} t} \cdot z_{j'}^{j} = \sum_{j'} v_{j'} \cdot e^{2\pi \boldsymbol{i} f_{j'} t} \sum_{j\in[2k]} \alpha_i(j) \cdot z_{j'}^{j} = 0.\end{aligned}$$

Hence for every $i \in [k]$,

$$-\alpha_i(l) \cdot f(t + l \cdot \Delta) = \sum_{j\in[2k]\setminus\{l\}} \alpha_i(j) \cdot f(t + j \cdot \Delta). \tag{3.4}$$

Let $A \in \mathbb{R}^{[k]\times[2k-1]}$ denote the matrix of the coefficients excluding the coordinate l, i.e.,

$$A = \begin{pmatrix} \alpha_1(0) & \cdots & \alpha_1(l-1) & \alpha_1(l+1) & \cdots & \alpha_1(2k-1) \\ \alpha_2(0) & \cdots & \alpha_2(l-1) & \alpha_2(l+1) & \cdots & \alpha_2(2k-1) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \alpha_k(0) & \cdots & \alpha_k(l-1) & \alpha_k(l+1) & \cdots & \alpha_k(2k-1) \end{pmatrix}.$$

For the $k \times k$ matrix $A \cdot A^*$, its entry (i, i') equals

$$\sum_{j\in[2k]\setminus\{l\}} \alpha_i(j) \cdot \overline{\alpha_{i'}(j)} = \langle \alpha_i, \alpha_{i'} \rangle - \alpha_i(l) \cdot \overline{\alpha_{i'}(l)} = 1_{i=i'} - \alpha_i(l) \cdot \overline{\alpha_{i'}(l)}.$$

Thus the eigenvalues of $A \cdot A^*$ are bounded by $1 + \sum_{i\in[k]} |\alpha_i(l)|^2$, which also bounds the eigenvalues of $A^* \cdot A$ by $1 + \sum_{i\in[k]} |\alpha_i(l)|^2$. From (3.4),

$$\begin{aligned}&\sum_{i\in[k]} |\alpha_i(l) \cdot f(t + l \cdot \Delta)|^2 \le \lambda_{\max}(A^* \cdot A) \cdot \sum_{j\in[2k]\setminus\{l\}} |f(t + j \cdot \Delta)|^2 \\ \Rightarrow \big(&\sum_{i\in[k]} |\alpha_i(l)|^2\big) \cdot |f(t + l \cdot \Delta)|^2 \le (1 + \sum_{i\in[k]} |\alpha_i(l)|^2) \cdot \sum_{j\in[2k]\setminus\{l\}} |f(t + j \cdot \Delta)|^2.\end{aligned}$$

Because $l = \arg\max_{j\in[2k]} \left\{ \sum_{i\in[k]} |\alpha_i(j)|^2 \right\}$ and $\alpha_1, \cdots, \alpha_k$ are unit vectors, $\sum_{i\in[k]} |\alpha_i(l)|^2 \geq \sum_{i=1}^k \|\alpha_i\|_2^2/2k \geq 1/2$. Therefore

$$|f(t + l\cdot\Delta)|^2 \leq 3 \sum_{j\in[2k]\setminus\{l\}} |f(t + j\cdot\Delta)|^2.$$

□

Corollary 3.2.3. *Given $f(x) = \sum_{j=1}^k v_j e^{2\pi \boldsymbol{i} f_j \cdot x}$, for any Δ and t,*

$$|f(t)|^2 \lesssim \sum_{i=1}^{2k} |f(t + i\Delta)|^2 + \sum_{i=1}^{2k} |f(t - i\Delta)|^2.$$

Next we finish the proof of Lemma 3.2.1.

Proof of Lemma 3.2.1. We assume $t = 1 - \epsilon$ and integrate Δ from 0 to $\epsilon/2k$:

$$\begin{aligned}
\varepsilon/2k \cdot |f(t)|^2 &\lesssim \int_{\Delta=0}^{\varepsilon/2k} \sum_{i=1}^{2k} |f(t + i\Delta)|^2 + \sum_{i=1}^{2k} |f(t - i\Delta)|^2 \mathrm{d}\Delta \\
&= \sum_{i\in[1,\ldots,2k]} \int_{\Delta=0}^{\varepsilon/2k} |f(t + i\Delta)|^2 + |f(t - i\Delta)|^2 \mathrm{d}\Delta \\
&\lesssim \sum_{i\in[1,\ldots,2k]} \frac{1}{i} \cdot \int_{\Delta'=0}^{\varepsilon\cdot i/2k} |f(t + \Delta')|^2 \mathrm{d}\Delta' + \sum_{i\in[1,\ldots,2k]} \frac{1}{i} \cdot \int_{\Delta'=0}^{\varepsilon\cdot i/2k} |f(t - \Delta')|^2 \mathrm{d}\Delta' \\
&\lesssim \sum_{i\in[1,\ldots,2k]} \frac{1}{i} \cdot \int_{\Delta'=-\varepsilon}^{\varepsilon} |f(t + \Delta')|^2 \mathrm{d}\Delta' \\
&\lesssim \log k \cdot \int_{x=-1}^{1} |f(x)|^2 \mathrm{d}x.
\end{aligned}$$

From all discussion above, we have $|f(1-\varepsilon)|^2 \lesssim \frac{k\log k}{\varepsilon} \cdot \mathop{\mathbb{E}}_{x\in[-1,1]}[|f(x)|^2]$. □

Next we finish the proof of Theorem 3.0.2.

Proof of Theorem 3.0.2. We bound

$$\begin{aligned}
\kappa &= \mathop{\mathbb{E}}_{x\in[-1,1]}[\sup_{f\in\mathcal{F}}\frac{|f(x)|^2}{\|f\|_2^2}] \\
&= \frac{1}{2}\int_{x=-1}^{1}\sup_{f\in\mathcal{F}}\frac{|f(x)|^2}{\|f\|_2^2}\mathrm{d}x \\
&\lesssim \int_{x=-1+\varepsilon}^{1-\varepsilon}\sup_{f\in\mathcal{F}}\frac{|f(x)|^2}{\|f\|_2^2}\mathrm{d}x + \varepsilon\cdot k^4\log^3 k && \text{from Theorem 3.0.1} \\
&\lesssim \int_{x=-1+\varepsilon}^{1-\varepsilon}\frac{k\log k}{1-|x|}\mathrm{d}x + \varepsilon\cdot k^4\log^3 k && \text{from Lemma 3.2.1} \\
&\lesssim k\log k\cdot\log\frac{1}{\varepsilon} + \varepsilon\cdot k^4\log^3 k \lesssim k\log^2 k
\end{aligned}$$

by choosing $\varepsilon = \frac{1}{k^3\log k}$. □

We show the distribution $D_{\mathcal{F}}$.

Proof of Corollary 3.0.3. From Claim 2.1.1 in Chapter 2, there exists a constant $c = \Theta(1)$ such that the distribution

$$D_{\mathcal{F}}(x) = \begin{cases} \frac{c}{(1-|x|)\log k}, & \text{for } |x| \le 1 - \frac{1}{k^3\log^2 k} \\ c\cdot k^3\log k, & \text{for } |x| > 1 - \frac{1}{k^3\log^2 k} \end{cases}$$

guarantees for any $f(x) = \sum_{j=1}^{k} v_j e^{2\pi \mathbf{i} f_j x}$, $\quad |f(x)|^2 \cdot \frac{D(x)}{D_{\mathcal{F}}(x)} = O(k\log^2 k)\cdot\|f\|_D^2 \quad \forall x\in[-1,1]$.

□

3.3 Polynomials and Fourier Sparse Signals

We prove that for any $\epsilon > 0$ and degree $k-1$ polynomial $p(x)$, there exists a signal $f \in \mathcal{F}$ such that $|f(x) - p(x)| \le \epsilon$ for all $x \in [-1,1]$. Then we show that there exists a degree $k-1$ polynomial $p(x)$ with $\frac{|p(1)|^2}{\|p\|_2^2} = k^2$, which complements the lower bound in Theorem 3.0.1.

Theorem 3.3.1. *For any degree $k-1$ polynomial $p(x) = \sum_{j=0}^{k-1} c_j x^j$ and $\epsilon > 0$, there always exists $\tau > 0$ and $f(x) = \sum_{j=0}^{k-1} a_j e^{2\pi \boldsymbol{i} \cdot (j\tau) \cdot x}$ such that*

$$\forall x \in [-1, 1], |f(x) - p(x)| \leq \epsilon.$$

At the same time, we use the Legendre polynomials to show the lower bound.

Theorem 3.3.2. *For any k, there exists a degree $k-1$ polynomial $p(x)$ such that*

$$\frac{|p(1)|^2}{\|p\|_2^2} = k^2.$$

At the same time, for any degree $k-1$ polynomial q and $x \in [-1, 1]$, $\frac{|q(x)|^2}{\|q\|_2^2} \leq k^2$.

We first prove Theorem 3.3.1 in Section 3.3.1 using the Taylor expansion. Then we review a few facts about the Legendre polynomials and prove Theorem 3.3.2 in Section 3.3.2.

3.3.1 A Reduction from Polynomials to Fourier Sparse Signals

We prove Theorem 3.3.1 in this section. We first consider the Taylor expansion of a signal $f(x) = \sum_{j=0}^{k-1} a_j e^{2\pi \boldsymbol{i} \cdot (j\tau) \cdot x}$ with a small frequency gap τ:

$$\begin{aligned} f(x) &= \sum_{j=0}^{k-1} a_j \sum_{l=0}^{\infty} \frac{(2\pi \boldsymbol{i} \cdot (j\tau) \cdot x)^l}{l!} \\ &= \sum_{l=0}^{\infty} \left(\frac{(2\pi \boldsymbol{i} \cdot \tau)^l}{l!} \cdot \sum_{j=0}^{k-1} a_j \cdot j^l \right) x^l. \end{aligned}$$

Given τ and a polynomial $p(x) = \sum_{l=0}^{k-1} c_l \cdot x^l$, we consider how to match their coefficients

$$\frac{(2\pi \boldsymbol{i} \cdot \tau)^l}{l!} \cdot \sum_{j=0}^{k-1} a_j \cdot j^l = c_l \text{ for every } l = 0, \ldots, k-1.$$

This is equivalent to

$$\sum_{j=0}^{k-1} j^l \cdot a_j = \frac{c_l \cdot l!}{(2\pi \boldsymbol{i} \cdot \tau)^l} \text{ for every } l = 0, \ldots, k-1.$$

Let A denote the $k \times k$ Vandermonde matrix $\left(j^l\right)_{l,j}$. Since $\det(A) = \prod_{i<j}(j-i) \neq 0$,

$$(a_0, \cdots, a_{k-1})^\top = A^{-1} \cdot \left(\frac{c_l \cdot l!}{(2\pi \boldsymbol{i} \cdot \tau)^l} \Big|_{l=0,\cdots,k-1} \right)^\top$$

satisfying the above equations.

Then we use the property of the Vandermonde matrix A to bound $a_j \leq \frac{\max_l\{c_l\} \cdot k^{k^2}}{\tau^{k-1}}$. $\det(A) = \prod_{i<j}(j-i) \geq 1$. On the other hand, $\lambda_{\max}(A) \leq k \cdot (k-1)^{k-1} \leq k^k$. Thus $\lambda_{\min}(A) \geq \frac{\det(A)}{\lambda_{\max}(A)^{k-1}} \geq k^{-k(k-1)}$.

From all discussion above, we have

$$\|(a_0, \cdots, a_{k-1})\|_2 \leq \lambda_{\min}(A)^{-1} \cdot \left\| \left(\frac{c_l \cdot l!}{(2\pi \boldsymbol{i} \cdot \tau)^l} \Big|_{l=0,\cdots,k-1} \right) \right\|_2.$$

This indicates

$$\max_i\{a_i\} \leq k^{k(k-1)} \frac{k^k \cdot \max_l\{c_l\}}{\tau^{k-1}} \leq \frac{\max_l\{c_l\} \cdot k^{k^2}}{\tau^{k-1}}.$$

Given k and the point-wise error ϵ, we set $\tau = \frac{\epsilon \cdot k^{-2k^2}}{\max_j\{c_j\}}$ such that the tail in the Taylor expansion is less than ϵ for $x \in [-1, 1]$:

$$\begin{aligned}
\sum_{l=k}^{\infty} \left(\frac{(2\pi \cdot \tau)^l}{l!} \cdot \sum_{j=0}^{k-1} a_j \cdot j^l \right) x^l &\leq \sum_{l=k}^{\infty} \frac{(2\pi \cdot \tau)^l}{l!} \cdot k \max_j\{a_j\} \cdot (k-1)^l \\
&\leq \sum_{l=k}^{\infty} k(2\pi \cdot \tau \cdot k)^l / l! \cdot \frac{\max_j\{c_j\} \cdot k^{k^2}}{\tau^{k-1}} \\
&\leq \sum_{l=k}^{\infty} k(2\pi \cdot k)^l / l! \cdot \max_j\{c_j\} \cdot k^{k^2} \cdot \tau^{l-k+1} \leq \epsilon.
\end{aligned}$$

3.3.2 Legendre Polynomials and a Lower Bound

We provide an brief introduction to Legendre polynomials (please see [Dun10] for a complete introduction).

Definition 3.3.3. *Let $L_n(x)$ denote the Legendre polynomials of degree n, the solution to Legendre's differential equation:*

$$\frac{\mathrm{d}}{\mathrm{d}x}\left[(1-x^2)\frac{\mathrm{d}}{\mathrm{d}x}L_n(x)\right] + n(n+1)L_n(x) = 0 \tag{3.5}$$

We will the following two facts about the Legendre polynomials in this work.

Fact 3.3.4. *$L_n(1) = 1$ for any $n \geq 0$ in the Legendre polynomials.*

Fact 3.3.5. *The Legendre polynomials constitute an orthogonal basis with respect to the inner product on interval $[-1, 1]$:*

$$\int_{-1}^{1} L_m(x)L_n(x)\mathrm{d}x = \frac{2}{2n+1}\delta_{mn}$$

where δ_{mn} denotes the Kronecker delta, i.e., it equals to 1 if $m = n$ and to 0 otherwise.

For any polynomial $P(x)$ of degree at most d with complex coefficients, there exists a set of coefficients from the above properties such that

$$P(x) = \sum_{i=0}^{d} \alpha_i \cdot L_i(x), \text{ where } \alpha_i \in \mathbb{C}, \forall i \in \{0, 1, 2, \cdots, d\}.$$

Now we finish the proof of Theorem 3.3.2.

Proof of Theorem 3.3.2. We first consider the upper bound. First observe that it is enough to prove $\frac{|q(1)|^2}{\|q\|_2^2} \leq k^2$ for any degree $k-1$ polynomial q due to the definition $\|q\|_2^2 = \mathop{\mathbb{E}}_{x\sim[-1,1]}\left[|q(x)|^2\right]$. For any q, let $x^* = \arg\max_{x\in[-1,1]} |q(x)|^2$. Then

$$\frac{|q(x^*)|^2}{\mathop{\mathbb{E}}_{x\sim[-1,x^*]}\left[|q(x)|^2\right]} \leq k^2 \text{ and } \frac{|q(x^*)|^2}{\mathop{\mathbb{E}}_{x\sim[x^*,1]}\left[|q(x)|^2\right]} \leq k^2$$

imply $\frac{|q(x^*)|^2}{\|q\|_2^2} \leq k^2$.

For any degree $k-1$ polynomial q, let $q(x) = \sum_{j=0}^{k-1} \alpha_j L_j(x)$ be its expression in the Legendre polynomials. Thus $q(1) = \sum_{j=0}^{k-1} \alpha_j$ and $\|q\|_2^2 = \sum_{j=0}^{k-1} |\alpha_j|^2/(2j+1)$. From the Cauchy-Schwartz inequality,

$$|q(1)|^2 \leq \left(\sum_{j=0}^{k-1} |\alpha_j|\right)^2 \leq \left(\sum_{j=0}^{k-1} |\alpha_j|^2/(2j+1)\right) \cdot \left(\sum_{j=0}^{k-1} 2j+1\right) = k^2 \cdot \|q\|_2^2.$$

On the other hand, there exists q such that the above inequality is tight. □

Chapter 4

Learning Continuous Sparse Fourier Transforms

We follow the notation in the last chapter. Let F be the bandlimit of the frequencies, $[-T, T]$ be the interval of observations, and $\mathcal{F} = \left\{ f(t) = \sum_{j=1}^{k} v_j \cdot e^{2\pi \boldsymbol{i} f_j t} \middle| v_j \in \mathbb{C}, |f_j| \le F \right\}$ be the family of signals. We consider the uniform distribution over $[-T, T]$ and define $\|f\|_2 = \left(\mathop{\mathbb{E}}_{t \in [-T,T]} [|f(t)|^2] \right)^{1/2}$ to denote the energy of a signal f.

In this chapter, we study the sample complexity of learning a signal $f \in \mathcal{F}$ under noise. Our main result is an sample-efficient algorithm under bounded ℓ_2 noise.

Theorem 4.0.1. *For any $F > 0, T > 0, \varepsilon > 0$, there exists an algorithm that given any observation $y(t) = \sum_{j=1}^{k} v_j e^{2\pi \boldsymbol{i} f_j t} + g(t)$ with $|f_j| \le F$ for each j, takes $m = O(k^4 \log^3 k + k^2 \log^2 k \cdot \log \frac{FT}{\varepsilon})$ samples $t_1, \ldots, t_m$ and outputs $\widetilde{f}(t) = \sum_{j=1}^{k} \widetilde{v}_j e^{2\pi \boldsymbol{i} \widetilde{f}_j t}$ satisfying*

$$\|f - \widetilde{f}\|_2^2 \lesssim \|g\|_2^2 + \epsilon \|f\|_2^2 \text{ with probability } 0.99.$$

We show our algorithm in Algorithm 1. To prove the correctness of Algorithm 1, we first state the technical result in this chapter.

Lemma 4.0.2. *Let C be a universal constant and $N_f = \frac{\varepsilon}{T \cdot k^{Ck^2}} \cdot \mathbb{Z} \cap [-F, F]$ denote a net of frequencies given F and T. For any signal $f(t) = \sum_{j=1}^{k} v_j e^{2\pi \boldsymbol{i} \cdot f_j t}$, there exists a k-sparse signal*

$$f'(t) = \sum_{j=1}^{k} v'_j e^{2\pi \boldsymbol{i} f'_j(t)} \text{ satisfying } \|f - f'\|_2 \le \varepsilon \|f\|_2 \text{ with frequencies } f'_1, \ldots, f'_k \in N_f.$$

Algorithm 1 Recover k-sparse FT

procedure SPARSEFT(y, F, T, ε)
  $m \leftarrow O(k^4 \log^3 k + k^2 \log^2 k \log \frac{FT}{\varepsilon})$
  Sample $t_1, \ldots, t_m$ from $D_{\mathcal{F}}$ where $D_{\mathcal{F}}$ is the distribution provided in Corollary 3.0.3.
  Set the corresponding weights $(w_1, \ldots, w_m)$ and $S = (t_1, \ldots, t_m)$
  Query $y(t_1), \ldots, y(t_m)$ from the observation y
  $N_f \leftarrow \frac{\varepsilon}{T \cdot k^{Ck^2}} \cdot \mathbb{Z} \cap [-F, F]$ for a constant C
  for all possible k frequencies $f'_1, \ldots, f'_k$ in N_f **do**
    Find $h(t)$ in span$\{e^{2\pi \boldsymbol{i} \cdot f'_1 t}, \ldots, e^{2\pi \boldsymbol{i} \cdot f'_k t}\}$ minimizing $\|h - y\|_{S,w}$
    Update $\widetilde{f} = h$ if $\|h - y\|_{S,w} \leq \|\widetilde{f} - y\|_{S,w}$
  end for
  Return $\widetilde{f}$.
end procedure

We prove Theorem 4.0.1 here then finish the proof of Lemma 4.0.2 in the rest of this chapter.

Proof of Theorem 4.0.1. We use Lemma 4.0.2 to rewrite $y = f + g = f' + g'$ where f' has frequencies in N_f and $g' = g + f - f'$ with $\|g'\|_2 \leq \|g\|_2 + \varepsilon \|f\|_2$. Our goal is to recover f'.

We construct a δ-net with $\delta = 0.05$ for

$$\left\{ h(t) = \sum_{j=1}^{2k} v_j e^{2\pi \boldsymbol{i} \cdot \widehat{h}_j t} \middle| \|h\|_2 = 1, \widehat{h}_j \in N_f \right\}.$$

We first pick $2k$ frequencies $\widehat{h}_1, \ldots, \widehat{h}_{2k}$ in N_f then construct a δ-net (ℓ_2-norm) on the linear subspace span$\{e^{2\pi \boldsymbol{i} \widehat{h}_1 t}, \ldots, e^{2\pi \boldsymbol{i} \widehat{h}_{2k} t}\}$. Hence the size of the δ-net is

$$\binom{\frac{4FT \cdot k^{Ck^2}}{\varepsilon}}{2k} \cdot (12/\delta)^{2k} \leq (\frac{4FT \cdot k^{Ck^2}}{\varepsilon \cdot \delta})^{3k}.$$

Now we consider the number of random samples from $D_{\mathcal{F}}$ to estimate signals in the δ-net. Based on the condition number of $D_{\mathcal{F}}$ in Theorem 3.0.2 and the Chernoff bound of Corollary 2.2.2, a union bound over the δ-net indicates

$$m = O\left(\frac{k \log^2 k}{\delta^2} \cdot \log |\text{net}| \right) = O\left(\frac{k \log^2 k}{\delta^2} \cdot (k^3 \log k + k \log \frac{FT}{\varepsilon \delta}) \right)$$

random samples from $D_{\mathcal{F}}$ would guarantee that for any signal h in the net, $\|h\|_{S,w}^2 = (1 \pm \delta)\|h\|_2^2$. From the property of the net, we obtain

$$\text{for any } h(t) = \sum_{j=1}^{2k} v_j e^{2\pi \boldsymbol{i} \cdot \widehat{h}_j t} \text{ with } \widehat{h}_j \in N_f, \quad \|h\|_{S,w}^2 = (1 \pm 2\delta)\|h\|_2^2.$$

Finally, we bound $\|f - \widetilde{f}\|_2$ as follows. The expectation of $\|f - \widetilde{f}\|_2$ over the random samples $S = (t_1, \ldots, t_m)$ is at most

$$\begin{aligned}
\|f - f'\|_2 + \|f' - \widetilde{f}\|_2 &\le \|f - f'\|_2 + 1.1\|f' - \widetilde{f}\|_{S,w} \\
&\le \|f - f'\|_2 + 1.1(\|f' - y\|_{S,w} + \|y - \widetilde{f}\|_{S,w}) \\
&\le \|f - f'\|_2 + 1.1(\|g'\|_{S,w} + \|y - f'\|_{S,w}) \\
&\le \varepsilon\|f\|_2 + 2.2(\|g\|_2 + \varepsilon\|f\|_2).
\end{aligned}$$

From the Markov inequality, with probability 0.99, $\|f - \widetilde{f}\|_2 \lesssim \varepsilon\|f\|_2 + \|g\|_2$. □

We sketch the proof of Lemma 4.0.2 here. Given $f(t) = \sum_{j=1}^{k} v_j e^{2\pi \boldsymbol{i} \cdot f_j t}$ where the frequencies $f_1, \cdots, f_k$ could be arbitrarily close to each other, we first show how to shift one frequency f_j to f_j' while keep almost the same signal $f(t)$ on $[-T, T]$.

Lemma 4.0.3. *There is a universal constant $C_0 > 0$ such that for any $f(t) = \sum_{j=1}^{k} v_j e^{2\pi \boldsymbol{i} f_j t}$ and any frequency f_{k+1}, there always exists*

$$f'(t) = \sum_{j=1}^{k-1} v_j' e^{2\pi \boldsymbol{i} f_j t} + v_{k+1}' e^{2\pi \boldsymbol{i} f_{k+1} t}$$

with k coefficients $v_1', v_2', \cdots, v_{k-1}', v_{k+1}'$ satisfying

$$\|f' - f\|_2 \le k^{C_0 k^2} \cdot (|f_k - f_{k+1}|T) \cdot \|f\|_2$$

We defer the proof of Lemma 4.0.3 to Section 4.2. Then we separate the frequencies $f_1, \cdots, f_k$ in $f(t)$ by at least η.

Lemma 4.0.4. *Given F and η, let $N_f = \eta\mathbb{Z} \cap [-F, F]$. For any k frequencies $f_1 < f_2 < \cdots < f_k$ in $[-F, F]$, there exists k frequencies $f'_1, \cdots, f'_k$ such that $\min_{i\in[k-1]} \{f'_{i+1} - f'_i\} \geq \eta$ and for all $i \in [k]$, $|f'_i - f_i| \leq k\eta$.*

Proof. Given a frequences f, let $\pi(f)$ denote the first element f' in N_f satisfying $f' \geq f$. We define the new frequencies f'_i as follows: $f'_1 = \pi(f_1)$ and $f'_i = \max\{f'_{i-1} + \eta, \pi(f_i)\}$ for $i \in \{2, 3, \cdots, k\}$. □

We finish the proof of Lemma 4.0.2 using the above two lemmas.

Proof of Lemma 4.0.2. Let $\eta = \frac{\varepsilon}{5k^2T\cdot k^{Ck^2}}$ for $C = C_0 + 1$ where C_0 is the universal constant in Lemma 4.0.3. Using Lemma 4.0.4 on frequencies $f_1, \cdots, f_k$ of the signal f, we obtain k new frequencies $f'_1, \cdots, f'_k$ such that their gap is at least η and $\max_i |f_i - f'_i| \leq k\eta$. Next we use the hybrid argument to find the signal f'.

Let $f^{(0)}(t) = f(t)$. For $i = 1, \cdots, k$, we apply Lemma 4.0.3 to shift the frequency f_i to f'_i and obtain

$$f^{(i)}(t) = \sum_{j=i+1}^{k} v_j^{(i)} e^{2\pi \boldsymbol{i} f_j t} + \sum_{j=1}^{i} v_j^{(i)} e^{2\pi \boldsymbol{i} f'_j t} \text{ s.t. } \|f^{(i)}(t) - f^{(i-1)}(t)\|_2 \leq k^{C_0k^2}(|f_i - f'_i|T)\|f^{(i-1)}\|_2.$$

At the same time, we bound $\|f^{(0)}(t)\|_2$ by

$$\left(1 - k^{C_0k^2}(k\eta T)\right)^i \|f^{(0)}(t)\|_2 \leq \|f^{(i)}(t)\|_2 \leq \left(1 + k^{C_0k^2}(k\eta T)\right)^i \|f^{(0)}(t)\|_2,$$

which is between $(1 \pm 0.1) \cdot \|f^{(0)}(t)\|_2$ for $\eta \leq \frac{1}{5k^2T} \cdot k^{-Ck^2}$ with $C = C_0 + 1$.

At last, we set $f'(t) = f^{(k)}(t)$ and bound the distance between $f'(t)$ and $f(t)$ by

$$\begin{aligned}
\|f^{(k)}(t) - f^{(0)}(t)\|_2 \leq & \sum_{i=1}^{k} \|f^{(i)}(t) - f^{(i-1)}(t)\|_2 && \text{by triangle inequality} \\
\leq & \sum_{i=1}^{k} k^{C_0 k^2} (|f_i - f_i'|T) \|f^{(i-1)}(t)\|_2 && \text{by Lemma 4.0.3} \\
\leq & \sum_{i=1}^{k} 2k^{C_0 k^2} (k\eta T) \|f^{(i-1)}(t)\|_2 && \text{by } \max_i |f_i - f_i'| \leq k\eta \\
\leq & 2k \cdot 2k^{C_0 k^2} (k\eta T) \|f(t)\|_2 \\
\leq & \epsilon \|f(t)\|_2
\end{aligned}$$

where the last inequality follows from our choice of $\eta = \frac{\varepsilon}{5k^2 T \cdot k^{Ck^2}}$. □

In the rest of this chapter, we review a few properties about the determinant of Gram matrices in Section 4.1 and finish the proof of Lemma 4.0.3 in Section 4.2.

4.1 Gram Matrices of Complex Exponentials

We provide an brief introduction to Gram matrices (please see [Haz01] for a complete introduction). We use $\langle x, y \rangle$ to denote the inner product between vector x and vector y.

Let $\vec{v}_1, \cdots, \vec{v}_n$ be n vectors in an inner product space and $\text{span}\{\vec{v}_1, \cdots, \vec{v}_n\}$ be the linear subspace spanned by these n vectors with coefficients in $\mathbb{C}$, i.e., $\left\{ \sum_{i \in [n]} \alpha_i \vec{v}_i | \forall i \in [n], \alpha_i \in \mathbb{C} \right\}$. The Gram matrix $\text{Gram}_{\vec{v}_1, \cdots, \vec{v}_n}$ of $\vec{v}_1, \cdots, \vec{v}_n$ is an $n \times n$ matrix defined as $\text{Gram}_{\vec{v}_1, \cdots, \vec{v}_n}(i, j) = \langle \vec{v}_i, \vec{v}_j \rangle$ for any $i \in [n]$ and $j \in [n]$.

Fact 4.1.1. $\det(\text{Gram}_{\vec{v}_1, \cdots, \vec{v}_n})$ *is the square of the volume of the parallelotope formed by* $\vec{v}_1, \cdots, \vec{v}_n$.

Let $\text{Gram}_{\vec{v}_1, \cdots, \vec{v}_{n-1}}$ be the Gram matrix of $\vec{v}_1, \cdots, \vec{v}_{n-1}$. Let $\vec{v}_n^{\parallel}$ be the projection of

v_n onto the linear subspace $\text{span}\{\vec{v}_1, \cdots, \vec{v}_{n-1}\}$ and $\vec{v}_n^{\perp} = \vec{v}_n - \vec{v}_n^{\parallel}$ be the part orthogonal to $\text{span}\{\vec{v}_1, \cdots, \vec{v}_{n-1}\}$. We use $\|\vec{v}\|$ to denote the length of $\vec{v}$ in the inner product space, which is $\sqrt{\langle \vec{v}, \vec{v} \rangle}$.

Claim 4.1.2.

$$\|\vec{v}_n^{\perp}\|^2 = \frac{\det(\text{Gram}_{\vec{v}_1, \cdots, \vec{v}_{n-1}})}{\det(\text{Gram}_{\vec{v}_1, \cdots, \vec{v}_n})}.$$

Proof.

$$\begin{aligned}\det(\text{Gram}_{\vec{v}_1, \cdots, \vec{v}_n}) =& \text{volume}^2(\vec{v}_1, \cdots, \vec{v}_n) \\ =& \text{volume}^2(\vec{v}_1, \cdots, \vec{v}_{n-1}) \cdot \|\vec{v}_n^{\perp}\|^2 = \det(\text{Gram}_{\vec{v}_1, \cdots, \vec{v}_{n-1}}) \cdot \|\vec{v}_n^{\perp}\|^2.\end{aligned}$$

□

We keep using the notation $e^{2\pi \boldsymbol{i} f_j t}$ to denote a vector from $[-T, T]$ to $\mathbb{C}$ and consider the inner product $\langle e^{2\pi \mathbf{i} f_i t}, e^{2\pi \mathbf{i} f_j t} \rangle_T = \frac{1}{2T} \int_{-T}^{T} e^{2\pi \mathbf{i}(f_i - f_j)t} \mathrm{d}t$ for complex exponential functions. We bound the determinant of the Gram matrices of $e^{2\pi \boldsymbol{i} f_1 t}, \ldots, e^{2\pi \boldsymbol{i} f_k t}$ as follows, which will extensively used in Section 4.2.

Lemma 4.1.3. *There exists a universal constant $\alpha > 0$ such that, for any $T > 0$ and real numbers $f_1, \cdots, f_k$, the $k \times k$ Gram matrix $\text{Gram}_{f_1, \cdots, f_k}$ of $e^{2\pi \mathbf{i} f_1 t}, e^{2\pi \mathbf{i} f_2 t}, \cdots, e^{2\pi \mathbf{i} f_k t}$, whose (i, j)-entry is*

$$\text{Gram}_{f_1, \cdots, f_k}(i, j) = \langle e^{2\pi \mathbf{i} f_i t}, e^{2\pi \mathbf{i} f_j t} \rangle_T,$$

has a determinant between

$$k^{-\alpha k^2} \prod_{i<j} \min((|f_i - f_j| T)^2, 1) \leq \det\left(\text{Gram}_{f_1, \cdots, f_k}\right) \leq k^{\alpha k^2} \prod_{i<j} \min((|f_i - f_j| T)^2, 1).$$

We prove this lemma in Section 4.1.1.

4.1.1 The Determinant of Gram Matrices of Complex Exponentials

Because we could rescale T to 1 and f_i to $f_i \cdot T$, we replace the interval $[-T, T]$ by $[-1, 1]$ and prove the following version: for real numbers $f_1, \ldots, f_k$, let $G_{f_1,\ldots,f_k}$ be the matrix whose (i, j)-entry is

$$\int_{-1}^{1} e^{2\pi \boldsymbol{i}(f_i - f_j)t} dt.$$

We plan to prove

$$\det(G_{f_1,\ldots,f_k}) = k^{O(k^2)} \prod_{i<j} \min(|f_i - f_j|^2, 1). \tag{4.1}$$

This indicates

$$\det(\mathrm{Gram}_{f_1,\cdots,f_k}) = 2^{-k} \cdot \det(G_{f_1,\ldots,f_k}) \in [k^{-\alpha k^2}, k^{\alpha k^2}] \cdot \prod_{i<j} \min((|f_i - f_j|T)^2, 1) \text{ for some } \alpha > 0.$$

First, we note by the Cauchy-Binet formula that the determinant in (4.1) is equal to

$$\int_{-1}^{1} \int_{-1}^{1} \cdots \int_{-1}^{1} \left|\det([e^{2\pi i f_i t_j}]_{i,j})\right|^2 dt_1 dt_2 \ldots dt_k. \tag{4.2}$$

We next need to consider the integrand in the special case when $\sum |f_i| \leq 1/8$.

Lemma 4.1.4. *If $f_i \in R$ and $t_j \in \mathbb{R}$, $\sum_i |f_i|(\max_i |t_i|) \leq 1/8$ then*

$$|\det([e^{2\pi i f_i t_j}]_{i,j})| = \Theta\left(\frac{(2\pi)^{\binom{k}{2}} \prod_{i<j} |t_i - t_j||f_i - f_j|}{1!2! \cdots k!}\right).$$

Proof. Firstly, by adding a constant to all the t_j we can make them non-negative. This multiplies the determinant by a root of unity, and at most doubles $\sum_i |f_i|(\max_i |t_i|)$.

By continuity, it suffices to consider the t_i to all be multiples of $1/N$ for some large integer N. By multiplying all the t_j by N and all f_i by $1/N$, we may assume that all of the t_j are non-negative integers with $t_1 \leq t_2 \leq \ldots \leq t_k$.

Let $z_i = \exp(2\pi i f_i)$. Then our determinant is

$$\det\left(\left[z_i^{t_j}\right]_{i,j}\right),$$

which is equal to the Vandermonde determinant times the Schur polynomial $s_\lambda(z_i)$ where λ is the partition $\lambda_j = t_j - (j-1)$.

Therefore, this determinant equals

$$\prod_{i<j}(z_i - z_j)s_\lambda(z_1, z_2, \ldots, z_k).$$

The absolute value of

$$\prod_{i<j}(z_i - z_j)$$

is approximately $\prod_{i<j}(2\pi i)(f_i - f_j)$, which has absolute value $(2\pi)^{\binom{k}{2}}\prod_{i<j}|f_i - f_j|$. We have left to evaluate the size of the Schur polynomial.

By standard results, s_λ is a polynomial in the z_i with non-negative coefficients, and all exponents at most $\max_j |t_j|$ in each variable. Therefore, the monomials with non-zero coefficients will all have real part at least $1/2$ and absolute value 1 when evaluated at the z_i. Therefore,

$$|s_\lambda(z_1, \ldots, z_k)| = \Theta(|s_\lambda(1, 1, \ldots, 1)|).$$

On the other hand, by the Weyl character formula

$$s_\lambda(1, 1, \ldots, 1) = \prod_{i<j}\frac{t_j - t_i}{j - i} = \frac{\prod_{i<j}|t_i - t_j|}{1!2!\ldots k!}.$$

This completes the proof. □

Next we prove our Theorem when the f have small total variation.

Lemma 4.1.5. *If there exists a f_0 so that $\sum |f_i - f_0| < 1/8$, then*

$$\det(G_{f_1,\ldots,f_k}) = \Theta\left(\frac{2^{3k(k-1)/2}\pi^{k(k-1)}\prod_{i<j}|f_i - f_j|^2}{(k!)^3\prod_{n=0}^{k-1}(2n)!}\right).$$

Proof. By translating the f_i we can assume that $f_0 = 0$.

By the above we have $\det(G_{f_1,\ldots,f_k})$ is

$$\Theta(\frac{(2\pi)^{k(k-1)}\prod_{i<j}|f_i - f_j|^2}{(1!2!\cdots k!)^2})\int_{-1}^{1}\cdots\int_{-1}^{1}\prod_{i<j}|t_i - t_j|^2 dt_1\ldots dt_k.$$

We note that by the Cauchy-Binet formula the latter term is the determinant of the matrix M with $M_{i,j} = \int_{-1}^{1} t^{i+j}dt$. This is the Graham matrix associated to the polynomials t^i for $0 \leq i \leq k-1$. Applying Graham-Schmidt (without the renormalization step) to this set yields the basis $P_n\alpha_n$ where $\alpha_n = \frac{2^n(n!)^2}{(2n)!}$ is the inverse of the leading term of P_n. This polynomial has norm $\alpha_n^2 2/(2n+1)$. Therefore, the integral over the t_i yields

$$\prod_{n=0}^{k-1}\frac{2^{n+1}(n!)^2}{(n+1)(2n)!}.$$

This completes the proof. □

Next we extend this result to the case that all the f are within $\mathrm{poly}(k)$ of each other.

Claim 4.1.6. *If there exists a f_0 so that $|f_i - f_0| = \mathrm{poly}(k)$ for all i, then*

$$\det(G_{f_1,\ldots,f_k}) = k^{O(k^2)}\prod_{i<j}\min(|f_i - f_j|^2, 1).$$

Proof. We begin by proving the lower bound. We note that for $0 < x < 1$,

$$\det(G_{f_1,\ldots,f_k}) \geq \int_{-x}^{x}\int_{-x}^{x}\cdots\int_{-1}^{1}\left|\det([e^{2\pi i f_i t_j}]_{i,j})\right|^2 dt_1 dt_2\ldots dt_k = x^k\det(G_{f_1/x,f_2/x,\ldots,f_k/k}).$$

Taking $x = 1/\mathrm{poly}(k)$, we may apply the above Lemma to compute the determinant on the right hand side, yielding an appropriate lower bound.

To prove the lower bound, we note that we can divide our f_i into clusters, $\mathcal{C}_i$, where for any i, j in the same cluster $|f_i - f_j| < 1/k$ and for i and j in different clusters $|f_i - f_j| \geq 1/k^2$. We then note as a property of Graham matrices that

$$\det(G_{f_1,\ldots,f_k}) \leq \prod_{\mathcal{C}_i} \det(G_{\{f_j \in \mathcal{C}_i\}}) = k^{O(k^2)} \prod_{i<j, \text{ in same cluster}} |f_i - f_j|^2 = k^{O(k^2)} \prod_{i<j} |f_i - f_j|^2.$$

This completes the proof. □

Finally, we are ready to prove our Theorem.

Proof. Let $I(t)$ be the indicator function of the interval $[-1, 1]$.

From Lemma 6.6 in [CKPS16], there is a function $h(t)$ so that for any function f that is a linear combination of at most k complex exponentials that $|h(t)f(t)|_2 = \Theta(|I(t)f(t)|_2)$ and so that $\hat{h}$ is supported on an interval of length $\text{poly}(k) < k^C$ about the origin.

Note that we can divide our f_i into clusters, $\mathcal{C}$, so that for i and j in a cluster $|f_i - f_j| < k^{C+1}$ and for i and j in different clusters $|f_i - f_j| > k^C$.

Let $\tilde{G}_{f_1,f_2,\ldots,f'_k}$ be the matrix with (i,j)-entry $\int_{\mathbb{R}} |h(t)|^2 e^{(2\pi i)(f_i - f_j)t} dt$.

We claim that for any $k' \leq k$ that

$$\det(\tilde{G}_{f_1,f_2,\ldots,f'_k}) = 2^{O(k')} \det(G_{f_1,f_2,\ldots,f'_k}).$$

This is because both are Graham determinants, one for the set of functions $I(t)\exp((2\pi i)f_j t)$ and the other for $h(t)\exp((2\pi i)f_j t)$. However since any linear combination of the former has L^2 norm a constant multiple of that the same linear combination of the latter, we have that

$$\tilde{G}_{f_1,f_2,\ldots,f'_k} = \Theta(G_{f_1,f_2,\ldots,f'_k})$$

as self-adjoint matrices. This implies the appropriate bound.

Therefore, we have that

$$\det(G_{f_1,\ldots,f_k}) = 2^{O(k)} \det(\tilde{G}_{f_1,\ldots,f_k}).$$

However, note that by the Fourier support of h that

$$\int_{\mathbb{R}} |h(t)|^2 e^{(2\pi i)(f_i - f_j)t} dt = 0$$

if $|f_i - f_j| > k^C$, which happens if i and j are in different clusters. Therefore $\tilde{G}$ is block diagonal and hence its determinant equals

$$\det(\tilde{G}_{f_1,\ldots,f_k}) = \prod_{\mathcal{C}} \det(\tilde{G}_{\{f_j \in \mathcal{C}_i\}}) = 2^{O(k)} \prod_{\mathcal{C}} \det(G_{\{f_j \in \mathcal{C}_i\}}).$$

However the Proposition above shows that

$$\prod_{\mathcal{C}} \det(G_{\{f_j \in \mathcal{C}_i\}}) = k^{O(k^2)} \prod_{i<j} \min(1, |f_i - f_j|^2).$$

This completes the proof. □

4.2 Shifting One Frequencies

We finish the proof of Lemma 4.0.3 in this section. We plan to shift f_k to f_{k+1} and prove that any vector in $\text{span}\{e^{2\pi \mathbf{i} f_1 t}, \cdots, e^{2\pi \mathbf{i} f_{k-1} t}, e^{2\pi \mathbf{i} f_k t}\}$ is close to some vector in the linear subspace $\text{span}\{e^{2\pi \mathbf{i} f_1 t}, \cdots, e^{2\pi \mathbf{i} f_{k-1} t}, e^{2\pi \mathbf{i} f_{k+1} t}\}$.

For convenience, we use $\vec{u}^{\parallel}$ to denote the projection of vector $e^{2\pi \mathbf{i} f_k t}$ to the linear subspace $U = \text{span}\{e^{2\pi \mathbf{i} f_1 t}, \cdots, e^{2\pi \mathbf{i} f_{k-1} t}\}$ and $\vec{w}^{\parallel}$ denote the projection of vector $e^{2\pi \mathbf{i} f_{k+1} t}$ to this linear subspace U. Let $\vec{u}^{\perp} = e^{2\pi \mathbf{i} f_k t} - \vec{u}^{\parallel}$ and $\vec{w}^{\perp} = e^{2\pi \mathbf{i} f_{k+1} t} - \vec{w}^{\parallel}$ be their orthogonal parts to U separately.

From the definition $e^{2\pi \mathbf{i} f_k t} = \vec{u}^{\parallel} + \vec{u}^{\perp}$ and $\vec{u}^{\parallel} \in U = \text{span}\{e^{2\pi \mathbf{i} f_1 t}, \cdots, e^{2\pi \mathbf{i} f_{k-1} t}\}$, we rewrite the linear combination

$$f(t) = \sum_{j=1}^{k} v_j e^{2\pi \mathbf{i} f_j t} = \sum_{j=1}^{k-1} \alpha_j e^{2\pi \mathbf{i} f_j t} + v_k \cdot \vec{u}^{\perp}$$

for some scalars $\alpha_1, \cdots, \alpha_{k-1}$.

We substitute $\vec{u}^\perp$ by $\vec{w}^\perp$ in the above linear combination and find a set of new coefficients. Let $\vec{w}^\perp = \vec{w}_1 + \vec{w}_2$ where $\vec{w}_1 = \frac{\langle \vec{u}^\perp, \vec{w}^\perp \rangle}{\|\vec{u}^\perp\|_2^2} \vec{u}^\perp$ is the projection of $\vec{w}^\perp$ to $\vec{u}^\perp$. Therefore $\vec{w}_2$ is the orthogonal part of the vector $e^{2\pi \mathbf{i} f_{k+1} t}$ to $V = \text{span}\{e^{2\pi \mathbf{i} f_1 t}, \cdots, e^{2\pi \mathbf{i} f_{k-1} t}, e^{2\pi \mathbf{i} f_k t}\}$. We use $\delta = \frac{\|\vec{w}_2\|_2}{\|\vec{w}^\perp\|_2}$ for convenience.

Notice that the $\min_\beta \frac{\|\vec{u}^\perp - \beta \cdot \vec{w}^\perp\|_2}{\|\vec{u}^\perp\|_2} = \delta$ and $\beta^* = \frac{\langle \vec{u}^\perp, \vec{w}^\perp \rangle}{\|\vec{w}^\perp\|_2^2}$ is the optimal choice. Therefore we set

$$f'(t) = \sum_{j=1}^{k-1} \beta_j e^{2\pi \mathbf{i} f_j t} + v_k \cdot \beta^* \cdot \vec{w}^\perp \in \text{span}\{e^{2\pi \mathbf{i} f_1 t}, \cdots, e^{2\pi \mathbf{i} f_{k-1} t}, e^{2\pi \mathbf{i} f_{k+1} t}\}$$

where the coefficients $\beta_1, \cdots, \beta_{k-1}$ guarantee that the projection of f' onto U is as same as the projection of f onto U. From the choice of β^* and the definition of f',

$$\|f(t) - f'(t)\|_2^2 = \delta^2 \cdot |v_k|^2 \cdot \|\vec{u}^\perp\|_2^2 \leq \delta^2 \cdot \|f(t)\|_2^2.$$

Eventually, we show an upper bound for δ^2 from Claim 4.1.2.

$$\begin{aligned}
\delta^2 &= \frac{\|\vec{w}_2\|_2^2}{\|\vec{w}^\perp\|_2^2} \\
&= \frac{\det(\text{Gram}_{e^{2\pi \mathbf{i} f_1 t}, \cdots, e^{2\pi \mathbf{i} f_{k-1} t}, e^{2\pi \mathbf{i} f_k t}, e^{2\pi \mathbf{i} f_{k+1} t}})}{\det(\text{Gram}_V)} \Big/ \frac{\det(\text{Gram}_{e^{2\pi \mathbf{i} f_1 t}, \cdots, e^{2\pi \mathbf{i} f_{k-1} t}, e^{2\pi \mathbf{i} f_{k+1} t}})}{\det(\text{Gram}_U)} \text{ by Claim 4.1.2} \\
&\leq k^{4\alpha k^2} \cdot \frac{\prod_{i=1}^{k+1} \prod_{\substack{j=1 \\ j \neq i}}^{k+1} \min(|f_i - f_j| T, 1)}{\prod_{i=1}^{k} \prod_{\substack{j=1 \\ j \neq i}}^{k} \min(|f_i - f_j| T, 1)} \cdot \frac{\prod_{i=1}^{k-1} \prod_{\substack{j=1 \\ j \neq i}}^{k-1} \min(|f_i - f_j| T, 1)}{\prod_{i=1}^{k-1} \prod_{\substack{j=1 \\ j \neq i}}^{k-1} \min(|f_i - f_j| T, 1) \cdot \prod_{i=1}^{k-1} \min(|f_i - f_{k+1}|^2 T^2, 1)} \\
&\quad \text{by Lemma 4.1.3} \\
&= k^{4\alpha k^2} |f_k - f_{k+1}|^2 T^2
\end{aligned}$$

Chapter 5

Fourier-clustered Signal Recovery

In this chapter, we reduce the recovery of signals $f(t)$, whose frequency representations $\widehat{f}$ are restricted to a small band $[-\Delta, \Delta]$, to the interpolation of low-degree polynomials. We first show that any such $f(t)$ could be approximated by low degree polynomials. In this section, we set $[-T, T]$ to be the interval of observations and use $\langle f, g\rangle = \frac{1}{2T}\int_{-T}^{T} f(t)\overline{g(t)}\mathrm{d}t$ denote the inner product of two signals f and g such that $\|f\|_2^2 = \langle f, f\rangle$.

Theorem 5.0.1. *For any $\Delta > 0$ and any $\varepsilon > 0$, let $f(t) = \sum_{j\in[k]} v_j e^{2\pi \boldsymbol{i} f_j t}$ where $|f_j| \le \Delta$ for each $j \in [k]$. There exists a polynomial $P(t)$ of degree at most*

$$d = O(T\Delta + k^3 \log k + k \log 1/\varepsilon)$$

such that

$$\|P(t) - f(t)\|_2^2 \le \varepsilon \|f\|_2^2.$$

In Chapter 3, Theorem 3.3.1 shows that any degree $k-1$ polynomial $P(t)$ could be approximated by a signal with k-sparse Fourier transform. Theorem 5.0.1 provides an approximation on the reverse direction. Notice that the dependency ΔT is necessary for a signal like $f(t) = \sin(2\pi \cdot \Delta t)$.

Next we provide an efficient algorithm that recovers polynomials under noise with an optimal sample complexity (up to a constant factor).

Theorem 5.0.2. *For any degree d polynomial $P(t)$ and an arbitrary noise function $g(t)$, there exists an algorithm that takes $O(d)$ samples from $x(t) = P(t) + g(t)$ over $[T, T]$ and*

reports a degree d polynomial $Q(t)$ in time $O(d^3)$ such that, with probability at least 99/100,

$$\|P(t) - Q(t)\|_2^2 \lesssim \|g(t)\|_2^2.$$

A direct corollary of the above two theorems indicates an efficient algorithm to recover signals $f(t)$ whose frequencies are restricted to a small band $[-\Delta, \Delta]$.

Corollary 5.0.3. *For any $k > 0$, $T > 0, \Delta > 0$, and $\varepsilon > 0$, there exist $d = O(T\Delta + k^3 \log k + k \log 1/\varepsilon)$ and an efficient algorithm that takes $O(d)$ samples from $y(t) = f(x) + g(t)$, where $f(t) = \sum_{j\in[k]} v_j e^{2\pi \boldsymbol{i} f_j t}$ with $|f_j| \le \Delta$ and g is an arbitrary noise function, and outputs a degree d polynomial $Q(t)$ in time $O(d^3)$ such that, with probability at least 99/100,*

$$\|f - Q\|_2^2 \lesssim \|g\|_2^2 + \varepsilon \|f\|_2^2.$$

In the rest of this chapter, we prove Theorem 5.0.1 in Section 5.1 and Theorem 5.0.2 in Section 5.2.

5.1 Band-limit Signals to Polynomials

We first prove a special case of Theorem 5.0.2 for k-Fourier-sparse signal with *a frequency gap* bounded away from zero. To prove this, we bound the coefficients $v_1, \cdots, v_k$ in $f(t) = \sum_{j=1}^{k} v_j e^{2\pi \boldsymbol{i} f_j t}$ by its energy $\|f\|_2^2$.

Lemma 5.1.1. *There exists a universal constant $c > 0$ such that for any $f(t) = \sum_{j=1}^{k} v_j e^{2\pi \boldsymbol{i} f_j t}$ with frequency gap $\eta = \min_{i\neq j} |f_i - f_j|$, we have $\|f(t)\|_2^2 \ge k^{-ck^2} \cdot \min\left((\eta T)^{2k}, 1\right) \cdot \sum_{j=1}^{k} |v_j|^2$.*

Proof. Let $\vec{v_i}$ denote the vector $e^{2\pi \boldsymbol{i} f_i t}$ and $V = \{\vec{v_1}, \cdots, \vec{v_k}\}$. Notice that $\|\vec{v_i}\|_2^2 = \langle \vec{v_i}, \vec{v_i} \rangle = 1$. For each $\vec{v_i}$, we define $\vec{v_i}^{\parallel}$ to be the projection of $\vec{v_i}$ into the linear subspace $\mathrm{span}\{V \setminus \vec{v_i}\} = \mathrm{span}\{\vec{v_1}, \cdots, \vec{v_{i-1}}, \vec{v_{i+1}}, \cdots, \vec{v_k}\}$ and $\vec{v_i}^{\perp} = \vec{v_i} - \vec{v_i}^{\parallel}$ which is orthogonal to $\mathrm{span}\{V \setminus \vec{v_i}\}$ by the definition.

Therefore from the orthogonality,

$$\|f(t)\|_2^2 \geq \max_{j\in[k]}\{|v_j|^2 \cdot \|\vec{v}_j^{\perp}\|_2^2\} \geq \frac{1}{k}\sum_{j=1}^{k}|v_j|^2 \cdot \|\vec{v}_j^{\perp}\|_2^2.$$

It is enough to estimate $\|\vec{v}_j^{\perp}\|_2^2$ from Claim 4.1.2:

$$\|\vec{v}_j^{\perp}\|_2^2 = \frac{\det(\mathrm{Gram}(V))}{\det(\mathrm{Gram}(V \setminus \vec{v}_i))} \geq k^{-2\alpha k^2}\prod_{j\neq i}\min\left((f_j - f_i)T, 1\right)^2 \geq k^{-2\alpha k^2}(\eta T)^{2k-2},$$

where we use Lemma 4.1.3 to lower bound it in the last step. □

We show that for signals with a frequency gap, its Taylor expansion is a good approximation.

Lemma 5.1.2 (Existence of low degree polynomial)**.** *Let* $f(t) = \sum_{j=1}^{k} v_j e^{2\pi \boldsymbol{i} f_j t}$*, where* $\forall j \in [k], |f_j| \leq \Delta$ *and* $\min_{i\neq j}|f_i - f_j| \geq \eta$*. There exists a polynomial* $Q(t)$ *of degree*

$$d = O\left(T\Delta + k\log 1/(\eta T) + k^2\log k + k\log(1/\varepsilon)\right)$$

such that,

$$\|Q(t) - f(t)\|_2^2 \leq \varepsilon\|f(t)\|_2^2. \tag{5.1}$$

Proof. For each frequency f_j, let $Q_j(t) = \sum_{k=0}^{d-1}\frac{(2\pi \boldsymbol{i} f_j t)^k}{k!}$ be the first d terms in the Taylor Expansion of $e^{2\pi \boldsymbol{i} f_j t}$. For any $t \in [-T, T]$, we know the difference between $Q_j(t)$ and $e^{2\pi \boldsymbol{i} f_j t}$ is at most

$$|Q_j(t) - e^{2\pi \boldsymbol{i} f_j t}| \leq |\frac{(2\pi \boldsymbol{i} f_j T)^d}{d!}| \leq (\frac{2\pi T\Delta \cdot e}{d})^d.$$

We define

$$Q(t) = \sum_{j=1}^{k} v_j Q_j(t)$$

and bound the distance between Q and f from the above estimation:

$$\begin{aligned}
\|Q(t)-f(t)\|_2^2 =& \frac{1}{2T}\int_{-T}^{T}|Q(t)-f(t)|^2\mathrm{d}t \\
=& \frac{1}{2T}\int_{-T}^{T}|\sum_{j=1}^{k}v_j(Q_j(t)-e^{2\pi \boldsymbol{i} f_j t})|^2\mathrm{d}t \\
\leq& 2k\sum_{j=1}^{k}\frac{1}{T}\int_{0}^{T}|v_j|^2\cdot|Q_j(t)-e^{2\pi \boldsymbol{i} f_j t}|^2\mathrm{d}t \qquad \text{by triangle inequality} \\
\leq& k\sum_{j=1}^{k}|v_j|^2\cdot(\frac{2\pi T\Delta\cdot e}{d})^{2d} \qquad \text{by Taylor expansion}
\end{aligned}$$

On the other hand, from Lemma 5.1.1, we know

$$\|f(t)\|_2^2 \geq (\eta T)^{2k}\cdot k^{-ck^2}\sum_j |v_j|^2.$$

Because $d = 10\cdot\pi e(T\Delta + k\log 1/(\eta T) + k^2\log k + k\log(1/\varepsilon))$ is large enough, we have $k(\frac{2\pi T\Delta\cdot e}{d})^{2d} \leq \varepsilon(\eta T)^{2k}\cdot k^{-ck^2}$, which indicates that $\|Q(t)-f(t)\|_2^2 \leq \varepsilon\|f\|_2^2$ from all discussion above. □

Proof of Theorem 5.0.1. We first use Lemma 4.0.2 on f to obtain f' with a frequency gap $\eta \geq \frac{\varepsilon}{T\cdot k^{Ck^2}}$ satisfying $\|f-f'\|_2 \leq \varepsilon\|f\|_2$. Then we use Lemma 5.1.2 on f' to obtain a degree $d = O\left(T\Delta + k\log 1/(\eta T) + k^2\log k + k\log(1/\varepsilon)\right) = O\left(T\Delta + k^3\log k + k\log(1/\varepsilon)\right)$ polynomial Q satisfying $\|Q-f'\|_2 \leq \varepsilon\|f'\|_2$. Hence

$$\|Q-f\|_2 \leq \|f-f'\|_2 + \|f'-Q\|_2 \leq 3\varepsilon\|f\|_2.$$

□

5.2 Robust Polynomial Interpolation

We show how to learn a degree-d polynomial P with $n = O(d)$ samples and prove Theorem 5.0.2 in this section. By define $\widetilde{f}(t) = f(t/T)$, we rescale the interval from $[-T, T]$ to $[-1, 1]$ and use $\|f\|_2^2 = \frac{1}{2}\int_{-1}^{1} |f(t)|^2 \mathrm{d}t$.

Lemma 5.2.1. *Let $d \in \mathbb{N}$ and $\epsilon \in \mathbb{R}^+$, there exists an efficient algorithm to compute a partition of $[-1, 1]$ to $n = O(d/\epsilon)$ intervals $I_1, \cdots, I_n$ such that for any degree d polynomial $P(t) : \mathbb{R} \to \mathbb{C}$ and any n points $t_1, \cdots, t_n$ in the intervals $I_1, \cdots, I_n$ respectively, the function $Q(t)$ defined by*

$$Q(t) = P(t_j) \quad \text{if} \quad t \in I_j$$

approximates P by

$$\|Q - P\|_2 \leq \epsilon \|P\|_2. \tag{5.2}$$

One direct corollary from the above lemma is that observing $n = O(d/\epsilon)$ points each from $I_1, \cdots, I_n$ provides a good approximation for all degree d polynomials. Recall that given a sequence $S = (t_1, \cdots, t_m)$ and corresponding weights $(w_1, \cdots, w_m)$, $\|f\|_{S,w} = (\sum_{i=1}^{m} w_i \cdot |f(t_i)|^2)^{1/2}$.

Corollary 5.2.2. *Let $I_1, \cdots, I_n$ be the intervals in the above lemma and $w_j = |I_j|/2$ for each $j \in [n]$. For any $t_1, \cdots, t_n$ in the intervals $I_1, \cdots, I_n$ respectively, let $S = (t_1, \cdots, t_n)$ with weights $(w_1, \cdots, w_n)$. Then for any degree d polynomial P, we have*

$$\|P\|_{S,w} \in \left[(1-\epsilon)\|P\|_2, (1+\epsilon)\|P\|_2\right].$$

We first prove one property of low degree polynomials from the Legendre basis.

Lemma 5.2.3. *For any degree d polynomial $P(t) : \mathbb{R} \to \mathbb{C}$ with derivative $P'(t)$, we have,*

$$\int_{-1}^{1} (1-t^2)|P'(t)|^2 \mathrm{d}t \leq 2d^2 \int_{-1}^{1} |P(t)|^2 \mathrm{d}t. \tag{5.3}$$

Proof. Given a degree d polynomial $P(x)$, we rewrite $P(x)$ as a linear combination of the Legendre polynomials:

$$P(x) = \sum_{i=0}^{d} \alpha_i L_i(x).$$

We use $F_i(x) = (1-x^2)L_i'(x)$ for convenience. From the definition of the Legendre polynomials in the Equation (3.5), $F_i'(x) = -i(i+1) \cdot L_i(x)$ and $F_i''(x) = -i(i+1) \cdot L_i'(x)$.

Hence we have

$$\begin{aligned}
\int_1^{-1} (1-x^2)|P'(x)|^2 \mathrm{d}x = & \int_1^{-1} (1-x^2) P'(x) \cdot \overline{P'}(x) \mathrm{d}x \\
= & \int_1^{-1} \left(\sum_{i\in[d]} \alpha_i F_i(x) \right) \cdot \left(\sum_{i\in[d]} \overline{\alpha_i} \frac{-F_i''(x)}{i(i+1)} \right) \mathrm{d}x \\
= & \left(\sum_{i\in[d]} \alpha_i F_i(x) \right) \cdot \left(\sum_{i\in[d]} \overline{\alpha_i} \frac{-F_i'(x)}{i(i+1)} \right) \Bigg|_{-1}^{1} \\
+ & \int_1^{-1} \left(\sum_{i\in[d]} \alpha_i F_i'(x) \right) \cdot \left(\sum_{i\in[d]} \overline{\alpha_i} \frac{F_i'(x)}{i(i+1)} \right) \mathrm{d}x \\
= & \int_1^{-1} \left(\sum_{i\in[d]} \alpha_i \cdot i(i+1) \cdot L_i(x) \right) \cdot \left(\sum_{i\in[d]} \overline{\alpha_i} \frac{i(i+1) \cdot L_i(x)}{i(i+1)} \right) \mathrm{d}x \\
= & \sum_{i\in[d]} |\alpha_i|^2 i(i+1) \|L_i\|_2^2 \\
\leq & \; d(d+1)\|P\|_2^2
\end{aligned}$$

□

Proof of Lemma 5.2.1. We set $m = 10d/\epsilon$ and show a partition of $[-1,1]$ into $n \leq 20m$ intervals. We define $g(t) = \frac{\sqrt{1-t^2}}{m}$ and $y_0 = 0$. Then we choose $y_i = y_{i-1} + g(y_{i-1})$ for $i \in \mathbb{N}^+$. Let l be the first index of y such that $y_l \geq 1 - \frac{9}{m^2}$. We show $l \lesssim m$.

Let j_k be the first index in the sequence such that $y_{j_k} \geq 1 - 2^{-k}$. Notice that

$$j_2 \leq \frac{3/4}{\frac{\sqrt{1-(3/4)^2}}{m}} \leq 1.5m$$

and

$$y_i - y_{i-1} = g(y_{i-1}) = \frac{\sqrt{1-y_{i-1}^2}}{m} \geq \frac{\sqrt{1-y_{i-1}}}{m}.$$

Then for all $k > 2$, we have

$$j_k - j_{k-1} \leq \frac{2^{-k}}{\frac{\sqrt{1-y_{(j_k-1)}}}{m}} \leq 2^{-k/2}m.$$

Therefore $j_k \leq \left(1.5 + (2^{-3/2} + \cdots 2^{-k/2})\right) m$ and $l \leq 10m$.

Because $y_{l-1} \leq 1 - \frac{9}{m^2}$, for any $j \in [l]$ and any $t \in [y_{i-1}, y_i]$, we have the following property:

$$\frac{1-t^2}{m^2} \geq \frac{1}{2} \cdot \frac{(1-y_{i-1}^2)}{m^2} = (y_i - y_{i-1})^2/2. \tag{5.4}$$

Now we set n and partition $[-1, 1]$ into $I_1, \cdots, I_n$ as follows:

1. $n = 2(l+1)$.
2. For $j \in [l]$, $I_{2j-1} = [y_{j-1}, y_j]$ and $I_{2j} = [-y_j, -y_{j-1}]$.
3. $I_{2l+1} = [y_l, 1]$ and $I_{2l+2} = [-1, -y_l]$.

For any $t_1, \cdots, t_n$ where $t_j \in I_j$ for each $j \in [n]$, we rewrite the LHS of (5.2) as follows:

$$\underbrace{\sum_{j=1}^{n-2} \int_{I_j} |P(t_j) - P(t)|^2 \, \mathrm{d}t}_{A} + \underbrace{\int_{I_{n-1}} |P(t_{n-1}) - P(t)|^2 \, \mathrm{d}t + \int_{I_n} |P(t_n) - P(t)|^2 \, \mathrm{d}t}_{B}. \tag{5.5}$$

For A in Equation (5.5), from the Cauchy-Schwarz inequality, we have

$$\sum_{j=1}^{n-2} \int_{I_j} |P(t_j) - P(t)|^2 \, \mathrm{d}t = \sum_{j=1}^{n-2} \int_{I_j} \left| \int_{t_j}^{t} P'(y) \mathrm{d}y \right|^2 \mathrm{d}t \leq \sum_{j=1}^{n-2} \int_{I_j} |t - t_j| \int_{t_j}^{t} |P'(y)|^2 \mathrm{d}y \mathrm{d}t.$$

Algorithm 2 RobustPolynomialLearningFixedInterval

procedure RobustPolynomialLearning(y, d)
  $\epsilon \leftarrow 1/20$.
  Let $I_1, \cdots, I_n$ be the intervals in Lemma 5.2.1 of parameters d and ϵ.
  Randomly choose $t_j \in I_j$ for every $j \in [n]$
  Define $S = \{t_1, \cdots, t_n\}$ with weight $w_1 = \frac{|I_1|}{2}, \cdots, w_n = \frac{|I_n|}{2}$.
  $Q(t) = \arg\min_{deg(Q)=d} \{\|Q - y\|_{S,w}\}$.
  Return $Q(t)$.
end procedure

Then we swap $\mathrm{d}t$ with $\mathrm{d}y$ and use Equation (5.4):

$$\sum_{j=1}^{n-2} \int_{I_j} |P'(y)|^2 \int_{t \notin (t_j, y)} |t - t_j| \mathrm{d}t \mathrm{d}y \leq \sum_{j=1}^{n-2} \int_{I_j} |P'(t)|^2 \cdot |I_j|^2 \mathrm{d}t \leq \sum_{j=1}^{n-2} \int_{I_j} |P'(t)|^2 \frac{2(1-t^2)}{m^2} \mathrm{d}t.$$

We use Lemma 5.2.3 to simplify it by

$$\sum_{j=1}^{n-2} \int_{I_j} |P(t_j) - P(t)|^2 \, \mathrm{d}t \leq \int_{-1}^{1} |P'(t)|^2 \frac{2(1-t^2)}{m^2} \mathrm{d}t \leq \frac{2d^2}{m^2} \int_{-1}^{1} |P(t)|^2 \mathrm{d}t.$$

For B in Equation (5.5), notice that $|I_{n-1}| = |I_n| = 1 - y_l \leq 9m^{-2}$ and for $j \in \{n-1, n\}$

$$|P(t) - P(t_j)|^2 \leq 4 \max_{t \in [-1,1]} |P(t)|^2 \leq 4(d+1)^2 \|P\|_2^2$$

from the properties of degree-d polynomials, i.e., Theorem 3.3.2. Therefore B in Equation (5.5) is upper bounded by $2 \cdot 4(d+1)^2 (9m^{-2}) \|P(t)\|_2^2$.

From all discussion above, $\|Q(t) - P(t)\|_2^2 \leq \frac{99d^2}{m^2} \leq \epsilon^2$. □

Now we use the above lemma to provide a faster learning algorithm for polynomials on interval $[-1, 1]$ with noise instead of using the ϵ-nets argument. We show it in Algorithm 2.

Lemma 5.2.4. *For any degree d polynomial $P(t)$ and an arbitrary function $g(t)$, Algorithm* RobustPolynomialLearningFixedInterval *takes $O(d)$ samples from $y(t) =$*

$P(t) + g(t)$ *over* $[-1, 1]$ *and reports a degree* d *polynomial* $Q(t)$ *in time* $O(d^3)$ *such that, with probability at least* $99/100$*,*

$$\|P(t) - Q(t)\|_2^2 \lesssim \|g(t)\|_2^2.$$

Proof. Notice that $n = O(d/\epsilon) = O(d)$ and find a degree d polynomial $Q(t)$ minimizing $\|y(t) - Q(t)\|_{S,w}$ is equivalent to calculate the pseudoinverse, which takes $O(d^3)$ time. It is enough to bound the distance between P and Q:

$$\begin{aligned}
&\|P - Q\|_2 \\
\leq &1.09\|P - Q\|_{S,w} && \text{by Corollary 5.2.2} \\
= &1.09\|y - g - Q\|_{S,w} && \text{by } y = P + g \\
\leq &1.09\|g\|_{S,w} + 1.09\|y - Q\|_{S,w} && \text{by triangle inequality} \\
\leq &1.09\|g\|_{S,w} + 1.09\|y - P\|_{S,w} && Q = \arg\min_{\text{degree-}d\ R} \|R - y\|_{S,w} \\
\leq &2.2\|g\|_{S,w}
\end{aligned}$$

Because $\mathop{\mathbb{E}}_S[\|g\|_{S,w}^2] = \|g\|_2^2$, we know that $\|P - Q\|_2 \lesssim \|g\|_2$ with probability $\geq .99$ by Markov's inequality. □

Chapter 6

Query and Active Learning of Linear Families

In Chapter 5, we show that $O(d)$ samples could robustly interpolate degree d polynomials under noisy observations. In this chapter, we generalize this result to any linear family under any distribution over its support and improve the guarantee of the output.

Let $\mathcal{F}$ be a linear family of dimension d and D be a distribution over the domain of $\mathcal{F}$. Recall that the worst-case condition number and the average condition number are

$$K_{\mathcal{F}} = \sup_{x \in \mathrm{supp}(D)} \sup_{f \in \mathcal{F}} \frac{|f(x)|^2}{\|f\|_2^2} \quad \text{and} \quad \kappa_{\mathcal{F}} = \mathop{\mathbb{E}}_{x \sim D} \left[\sup_{f \in \mathcal{F}} \frac{|f(x)|^2}{\|f\|_2^2} \right].$$

We first show that for any linear family $\mathcal{F}$ and any distribution D over the support of $\mathcal{F}$, the average condition number $\kappa_{\mathcal{F}} = d$. For convenience, let $\langle f, g \rangle = \mathop{\mathbb{E}}_{x \sim D}[f(x)\overline{g(x)}]$ denote the inner product under D and $\|f\|_D = (\mathop{\mathbb{E}}_{x \sim D}[|f(x)|^2])^{1/2}$ in this chapter.

Lemma 6.0.1. *For any linear family $\mathcal{F}$ of dimension d,*

$$\mathop{\mathbb{E}}_{x \sim D} \sup_{h \in \mathcal{F}: \|h\|_D = 1} |h(x)|^2 = d$$

such that $D_{\mathcal{F}}(x) = D(x) \cdot \sup_{h \in \mathcal{F}: \|h\|_D = 1} |h(x)|^2/d$ has a condition number $K_{D_{\mathcal{F}}} = d$. Moreover, there exists an efficient algorithm to sample x from $D_{\mathcal{F}}$ and compute its weight $\frac{D(x)}{D_{\mathcal{F}}(x)}$.

Based on this condition number d, we use the matrix Chernoff bound to show that $O(d \log d)$ i.i.d. samples from $D_{\mathcal{F}}$ suffice to learn $\mathcal{F}$.

However, this approach needs $\Omega(d \log d)$ samples due to a coupon-collector argument, because it only samples points from one distribution $D_{\mathcal{F}}$. Next we consider how to improve

it to $O(d)$ using linear size spectral sparsification [BSS12, LS15]. We need our sample points to both be sampled non-independently (to avoid coupon-collector issues) but still fairly randomly (so adversarial noise cannot predict it). A natural approach is to design a sequence of distributions $D_1, \cdots, D_m$ (m is not necessarily fixed) then sample $x_i \sim D_i$ and assign a weight w_i for x_i, where D_{i+1} could depend on the previous points $x_1, \cdots, x_i$.

Ideally each K_{D_i} would be $O(d)$, but we do not know how to produce such distributions while still getting linear sample spectral sparsification to guarantee $\|h\|_{S,w} \approx \|h\|_D$ for every $h \in \mathcal{F}$. Therefore we use a coefficient α_i to control every K_{D_i}, and set $w_i = \alpha_i \cdot \frac{D(x_i)}{D_i(x_i)}$ instead of $\frac{D(x_i)}{mD_i(x_i)}$.

Definition 6.0.2. *Given a linear family $\mathcal{F}$ and underlying distribution D, let P be a random sampling procedure that terminates in m iterations (m is not necessarily fixed) and provides a coefficient α_i and a distribution D_i to sample $x_i \sim D_i$ in every iteration $i \in [m]$.*

We say P is an ε-well-balanced sampling procedure *if it satisfies the following two properties:*

1. *Let the weight $w_i = \alpha_i \cdot \frac{D(x_i)}{D_i(x_i)}$. With probability $1 - 10^{-3}$,*

$$\sum_{i=1}^{m} w_i \cdot |h(x_i)|^2 \in \left[\frac{3}{4}, \frac{5}{4}\right] \cdot \|h\|_D^2 \quad \forall h \in \mathcal{F}.$$

2. *For a universal constant C, the coefficients always have $\sum_i \alpha_i \le \frac{5}{4}$ and $\alpha_i \cdot K_{D_i} \le \epsilon/C$.*

Intuitively, the first property says that the sampling procedure preserves the signal, and the second property says that the recovery algorithm does not blow up the noise on average. For such sampling procedures we consider the weighted ERM $\widetilde{f} \in \mathcal{F}$ minimizing $\sum_i w_i |\widetilde{f}(x_i) - y_i|^2$. We prove that $\widetilde{f}$ satisfies the desired guarantee:

Theorem 6.0.3. *Given a linear family $\mathcal{F}$, joint distribution (D, Y), and $\varepsilon > 0$, let P be an ε-well-balanced sampling procedure for $\mathcal{F}$ and D, and let $f = \arg\min_{h \in \mathcal{F}} \mathop{\mathbb{E}}_{(x,y)\sim(D,Y)}[|y - h(x)|^2]$ be the true risk minimizer. Then the weighted ERM $\widetilde{f}$ resulting from P satisfies*

$$\|f - \widetilde{f}\|_D^2 \leq \epsilon \cdot \mathop{\mathbb{E}}_{(x,y)\sim(D,Y)}[|y - f(x)|^2]$$

with 99% probability.

Well-balanced sampling procedures. We observe that two standard sampling procedures are well-balanced, so they yield agnostic recovery guarantees by Theorem 6.0.3. The simplest approach is to set each D_i to a fixed distribution D' and $\alpha_i = 1/m$ for all i. For $m = O(K_{D'} \log d + K_{D'}/\varepsilon)$, this gives an ε-well-balanced sampling procedure. These results appear in Section 6.3.

We get a stronger result of $m = O(d/\varepsilon)$ using the randomized BSS algorithm by Lee and Sun [LS15]. The [LS15] algorithm iteratively chooses points x_i from distributions D_i. A term considered in their analysis—the largest increment of eigenvalues—is equivalent to our K_{D_i}. By looking at the potential functions in their proof, we can extract coefficients α_i bounding $\alpha_i K_{D_i}$ in our setting. This lets us show that the algorithm is a well-balanced sampling procedure; we do so in Section 6.4.

Next we generalize the above result to active learning, where the algorithm receives a set of unlabelled points x_i from the *unknown* distribution D and chooses a subset of these points to receive their labels.

Theorem 6.0.4. *Consider any dimension d linear space $\mathcal{F}$ of functions from a domain G to $\mathbb{C}$. Let (D, Y) be a joint distribution over $G \times \mathbb{C}$ and $f = \arg\min_{h \in \mathcal{F}} \mathop{\mathbb{E}}_{(x,y)\sim(D,Y)}[|y - h(x)|^2]$.*

Let $K = \sup_{h \in \mathcal{F}: h \neq 0} \frac{\sup_{x \in G} |h(x)|^2}{\|h\|_D^2}$. For any $\varepsilon > 0$, there exists an efficient algorithm that

takes $O(K \log d + \frac{K}{\varepsilon})$ *unlabeled samples from* D *and requests* $O(\frac{d}{\varepsilon})$ *labels to output* $\widetilde{f}$ *satisfying*

$$\mathop{\mathbb{E}}_{x\sim D}[|\tilde{f}(x) - f(x)|^2] \le \varepsilon \cdot \mathop{\mathbb{E}}_{(x,y)\sim(D,Y)}[|y - f(x)|^2] \text{ with probability } \ge 0.99.$$

Finally we show two lower bounds on the sample complexity and query complexity that match our upper bounds.

	Lower bound	Upper bound
Query complexity	Theorem 6.6.1 $\Omega(\frac{d}{\varepsilon})$	Theorem 6.4.2 $O(\frac{d}{\varepsilon})$
Sample complexity	Theorem 6.6.4 $\Omega(K \log d + \frac{K}{\varepsilon})$	Theorem 6.3.3 $O(K \log d + \frac{K}{\varepsilon})$

Table 6.1: Lower bounds and upper bounds in different models

Organization. We organize the rest of this chapter as follows. In Section 6.1, we show the distribution $D_{\mathcal{F}}$ to prove Lemma 6.1.1. Then we discuss the ERM of well-balanced sampling procedures and prove Theorem 6.0.3 in Section 6.2. In Section 6.3 we analyze the number of samples required for sampling from an arbitrary distribution D' to be well-balanced. In Section 6.4 we show that [LS15] yields a well-balanced linear-sample procedure in the query setting or in fixed-design active learning. Next we show the active learning algorithm of Theorem 6.0.4 in Section 6.5. Finally, we prove the lower bounds on the query complexity and sample complexity in Table 6.1 in Section 6.6.

6.1 Condition Number of Linear families

We use the linearity of $\mathcal{F}$ to prove $\kappa = d$ and describe the distribution $D_{\mathcal{F}}$. Let $\{v_1, \ldots, v_d\}$ be any orthonormal basis of $\mathcal{F}$, where inner products are taken under the distribution D.

Lemma 6.1.1. *For any linear family* $\mathcal{F}$ *of dimension* d *and any distribution* D*,*

$$\mathop{\mathbb{E}}_{x\sim D} \sup_{h\in\mathcal{F}:\|h\|_D=1} |h(x)|^2 = d$$

Algorithm 3 SampleDF

1: **procedure** GENERATINGDF($\mathcal{F} = \text{span}\{v_1, \ldots, v_d\}, D$)
2: Sample $j \in [d]$ uniformly.
3: Sample x from the distribution $W_j(x) = |v_j(x)|^2$.
4: Set the weight of x to be $\frac{d}{\sum_{i=1}^{d} |v_i(x)|^2}$.
5: **end procedure**

such that $D_{\mathcal{F}}(x) = D(x) \cdot \sup_{h \in \mathcal{F}: \|h\|_D = 1} |h(x)|^2/d$ *has a condition number* $K_{D_{\mathcal{F}}} = d$. *Moreover, there exists an efficient algorithm to sample* x *from* $D_{\mathcal{F}}$ *and compute its weight* $\frac{D(x)}{D_{\mathcal{F}}(x)}$.

Proof. Given an orthonormal basis $v_1, \ldots, v_d$ of $\mathcal{F}$, for any $h \in \mathcal{F}$ with $\|h\|_D = 1$, there exists $c_1, \ldots, c_d$ such that $h(x) = c_i \cdot v_i(x)$. Then for any x in the domain, from the Cauchy-Schwartz inequality,

$$\sup_h \frac{|h(x)|^2}{\|h\|_D^2} = \sup_{c_1,\ldots,c_d} \frac{|\sum_{i\in[d]} c_i v_i(x)|^2}{\sum_{i\in[d]} |c_i|^2} = \frac{(\sum_{i\in[d]} |c_i|^2) \cdot (\sum_{i\in[d]} |v_i(x)|^2)}{\sum_{i\in[d]} |c_i|^2} = \sum_{i\in[d]} |v_i(x)|^2.$$

This is tight because there always exist $c_1 = \overline{v_1(x)}, c_2 = \overline{v_2(x)}, \ldots, c_d = \overline{v_d(x)}$ such that $|\sum_{i\in[d]} c_i v_i(x)|^2 = (\sum_{i\in[d]} |c_i|^2) \cdot (\sum_{i\in[d]} |v_i(x)|^2)$. Hence

$$\mathop{\mathbb{E}}_{x\sim D} \sup_{h\in\mathcal{F}: h\neq 0} \frac{|h(x)|^2}{\|h\|_D^2} = \mathop{\mathbb{E}}_{x\sim D} \Big[\sum_{i\in[d]} |v_i(x)|^2 \Big] = d.$$

By Claim 2.1.1, this indicates $K_{D_{\mathcal{F}}} = d$. At the same time, this calculation indicates

$$D_{\mathcal{F}}(x) = \frac{D(x) \cdot \sup_{\|h\|_D = 1} |h(x)|^2}{d} = \frac{D(x) \cdot \sum_{i\in[d]} |v_i(x)|^2}{d}.$$

We present our sampling procedure in Algorithm 3.

□

6.2 Recovery Guarantee for Well-Balanced Samples

In this section, we show for well-balanced sampling procedures (per Definition 6.0.2) that an appropriately weighted ERM approximates the true risk minimizer, and hence the true signal.

Definition 6.2.1. *Given a random sampling procedure P, and a joint distribution (D, Y), we define the weighted ERM resulting from P to be the result*

$$\widetilde{f} = \arg\min_{h \in \mathcal{F}} \left\{ \sum_{i=1}^{m} w_i \cdot |h(x_i) - y_i|^2 \right\}$$

after we use P to generate random points $x_i \sim D_i$ with $w_i = \alpha_i \cdot \frac{D(x_i)}{D_i(x_i)}$ for each i, and draw $y_i \sim (Y \mid x_i)$ for each point x_i.

For generality, we first consider points and labels from a joint distribution (D, Y) and use $f = \arg\min_{h \in \mathcal{F}} \mathbb{E}_{(x,y)\sim(D,Y)}[|y - h(x)|^2]$ to denote the truth.

Theorem 6.0.3. *Given a linear family $\mathcal{F}$, joint distribution (D, Y), and $\varepsilon > 0$, let P be an ε-well-balanced sampling procedure for $\mathcal{F}$ and D, and let $f = \arg\min_{h \in \mathcal{F}} \mathbb{E}_{(x,y)\sim(D,Y)}[|y - h(x)|^2]$ be the true risk minimizer. Then the weighted ERM $\widetilde{f}$ resulting from P satisfies*

$$\|f - \widetilde{f}\|_D^2 \leq \epsilon \cdot \mathbb{E}_{(x,y)\sim(D,Y)}[|y - f(x)|^2]$$

with 99% probability.

Next, we provide a corollary for specific kinds of noise. In the first case, we consider noise functions representing independently mean-zero noise at each position x such as i.i.d. Gaussian noise. Second, we consider arbitrary noise functions on the domain.

Corollary 6.2.2. *Given a linear family $\mathcal{F}$ and distribution D, let $y(x) = f(x) + g(x)$ for $f \in \mathcal{F}$ and g a randomized function. Let P be an ε-well-balanced sampling procedure for $\mathcal{F}$ and D. With 99% probability, the weighted ERM $\widetilde{f}$ resulting from P satisfies*

1. $\|\tilde{f} - f\|_D^2 \le \varepsilon \cdot \mathop{\mathbb{E}}_g[\|g\|_D^2]$, *when* $g(x)$ *is a random function from* G *to* $\mathbb{C}$ *where each* $g(x)$ *is an independent random variable with* $\mathop{\mathbb{E}}_g[g(x)] = 0$.

2. $\|\tilde{f} - f\|_D \le (1+\varepsilon) \cdot \|g\|_D$ *for any other noise function* g.

In the rest of this section, we prove Theorem 6.0.3 in Section 6.2.1 and Corollary 6.2.2 in Section 6.2.2.

6.2.1 Proof of Theorem 6.0.3

We introduce a few more notation in this proof. Given $\mathcal{F}$ and the measurement D, let $\{v_1, \ldots, v_d\}$ be a fixed orthonormal basis of $\mathcal{F}$, where inner products are taken under the distribution D, i.e., $\mathop{\mathbb{E}}_{x\sim D}[v_i(x) \cdot \overline{v_j(x)}] = 1_{i=j}$ for any $i, j \in [d]$. For any function $h \in \mathcal{F}$, let $\alpha(h)$ denote the coefficients $(\alpha(h)_1, \ldots, \alpha(h)_d)$ under the basis $(v_1, \ldots, v_d)$ such that $h = \sum_{i=1}^d \alpha(h)_i \cdot v_i$ and $\|\alpha(h)\|_2 = \|h\|_D$.

We characterize the first property in Definition 6.0.2 of *well-balanced sampling procedures* as bounding the eigenvalues of $A^* \cdot A$, where A is the $m \times d$ matrix defined as $A(i,j) = \sqrt{w_i} \cdot v_j(x_i)$.

Lemma 6.2.3. *For any* $\varepsilon > 0$, *given* $S = (x_1, \ldots, x_m)$ *and their weights* $(w_1, \ldots, w_m)$, *let* A *be the* $m \times d$ *matrix defined as* $A(i,j) = \sqrt{w_i} \cdot v_j(x_i)$. *Then*

$$\|h\|_{S,w}^2 \in [1 \pm \varepsilon] \cdot \|h\|_D^2 \quad \text{for every } h \in \mathcal{F}$$

if and only if the eigenvalues of A^*A *are in* $[1-\varepsilon, 1+\varepsilon]$.

Proof. Notice that

$$A \cdot \alpha(h) = \left(\sqrt{w_1} \cdot h(x_1), \ldots, \sqrt{w_m} \cdot h(x_m)\right). \tag{6.1}$$

Because

$$\|h\|_{S,w}^2 = \sum_{i=1}^m w_i |h(x_i)|^2 = \|A \cdot \alpha(h)\|_2^2 = \alpha(h)^* \cdot (A^* \cdot A) \cdot \alpha(h) \in [\lambda_{\min}(A^* \cdot A), \lambda_{\max}(A^* \cdot A)] \cdot \|h\|_D^2$$

and h is over the linear family $\mathcal{F}$, these two properties are equivalent.

□

Next we consider the calculation of the weighted ERM $\widetilde{f}$. Given the weights $(w_1, \cdots, w_m)$ on $(x_1, \ldots, x_m)$ and labels $(y_1, \ldots, y_m)$, let $\vec{y}_w$ denote the vector of weighted labels $(\sqrt{w_1} \cdot y_1, \ldots, \sqrt{w_m} \cdot y_m)$. From (6.1), the empirical distance $\|h-(y_1, \ldots, y_m)\|_{S,w}^2$ equals $\|A \cdot \alpha(h) - \vec{y}_w\|_2^2$ for any $h \in \mathcal{F}$. The function $\widetilde{f}$ minimizing $\|h - (y_1, \ldots, y_m)\|_{S,w} = \|A \cdot \alpha(h) - \vec{y}_w\|_2$ overall all $h \in \mathcal{F}$ is the pseudoinverse of A on $\vec{y}_w$, i.e.,

$$\alpha(\widetilde{f}) = (A^* \cdot A)^{-1} \cdot A^* \cdot \vec{y}_w \text{ and } \widetilde{f} = \sum_{i=1}^{d} \alpha(\widetilde{f})_i \cdot v_i.$$

Finally, we consider the distance between $f = \arg\min_{h \in \mathcal{F}} \big\{ \mathop{\mathbb{E}}_{(x,y)\sim(D,Y)} [|h(x) - y|^2] \big\}$ and $\widetilde{f}$. For convenience, let $\vec{f}_w = \big(\sqrt{w_1} \cdot f(x_1), \ldots, \sqrt{w_m} \cdot f(w_m)\big)$. Because $f \in \mathcal{F}$, $(A^* \cdot A)^{-1} \cdot A^* \cdot \vec{f}_w = \alpha(f)$. This implies

$$\|\widetilde{f} - f\|_D^2 = \|\alpha(\widetilde{f}) - \alpha(f)\|_2^2 = \|(A^* \cdot A)^{-1} \cdot A^* \cdot (\vec{y}_w - \vec{f}_w)\|_2^2.$$

We assume $\lambda\big((A^* \cdot A)^{-1}\big)$ is bounded and consider $\|A^* \cdot (\vec{y}_w - \vec{f}_w)\|_2^2$.

Lemma 6.2.4. *Let P be an random sampling procedure terminating in m iterations (m is not necessarily fixed) that in every iteration i, it provides a coefficient α_i and a distribution D_i to sample $x_i \sim D_i$. Let the weight $w_i = \alpha_i \cdot \frac{D(x_i)}{D_i(x_i)}$ and $A \in \mathbb{C}^{m \times d}$ denote the matrix $A(i,j) = \sqrt{w_i} \cdot v_j(x_i)$. Then for $f = \arg\min_{h \in \mathcal{F}} \mathop{\mathbb{E}}_{(x,y)\sim(D,Y)} [|y - h(x)|^2]$,*

$$\mathop{\mathbb{E}}_{P} \Big[\|A^*(\vec{y}_w - \vec{f}_{S,w})\|_2^2 \Big] \leq \sup_P \Big\{ \sum_{i=1}^{m} \alpha_i \Big\} \cdot \max_j \big\{ \alpha_j \cdot K_{D_j} \big\} \mathop{\mathbb{E}}_{(x,y)\sim(D,Y)} [|y - f(x)|^2],$$

where K_{D_i} is the condition number for samples from D_i: $K_{D_i} = \sup_x \Big\{ \frac{D(x)}{D_i(x)} \cdot \sup_{v \in \mathcal{F}} \big\{ \frac{|v(x)|^2}{\|v\|_2^2} \big\} \Big\}$.

Proof. For convenience, let g_j denote $y_j - f(x_j)$ and $\vec{g}_w \in \mathbb{C}^m$ denote the vector $\left(\sqrt{w_j} \cdot g_j|_{j=1,\ldots,m}\right) = \vec{y}_w - \vec{f}_{S,w}$ for $j \in [m]$ such that $A^* \cdot (\vec{y}_w - \vec{f}_{S,w}) = A^* \cdot \vec{g}_w$.

$$\begin{aligned}\mathbb{E}[\|A^* \cdot \vec{g}_w\|_2^2] &= \mathbb{E}\left[\sum_{i=1}^{d} \big(\sum_{j=1}^{m} A^*(i,j)\vec{g}_w(j)\big)^2\right] \\ &= \sum_{i=1}^{d} \mathbb{E}\left[\big(\sum_{j=1}^{m} w_j \overline{v_i(x_j)} \cdot g_j\big)^2\right] = \sum_{i=1}^{d} \mathbb{E}\left[\sum_{j=1}^{m} w_j^2 \cdot |v_i(x_j)|^2 \cdot |g_j|^2\right],\end{aligned}$$

where the last step uses the following fact

$$\mathop{\mathbb{E}}_{w_j \sim D_j}[w_j \overline{v_i(x_j)} \cdot g_j] = \mathop{\mathbb{E}}_{w_j \sim D_j}\Big[\alpha_j \cdot \frac{D(x_j)}{D_j(x_j)} \overline{v_i(x_j)} g_j\Big] = \alpha_j \cdot \mathop{\mathbb{E}}_{x_j \sim D, y_j \sim Y(x_j)}\big[\overline{v_i(x_j)}(y_j - f(x_j))\big] = 0.$$

We swap i and j:

$$\begin{aligned}\sum_{i=1}^{d} \mathbb{E}\left[\sum_{j=1}^{m} w_j^2 \cdot |v_i(x_j)|^2 \cdot |g_j|^2\right] &= \sum_{j=1}^{m} \mathbb{E}\left[\sum_{i=1}^{d} w_j |v_i(x_j)|^2 \cdot w_j |g_j|^2\right] \\ &\le \sum_{j=1}^{m} \sup_{x_j}\left\{w_j \sum_{i=1}^{d} |v_i(x_j)|^2\right\} \cdot \mathbb{E}\left[w_j \cdot |g_j|^2\right].\end{aligned}$$

For $\mathbb{E}\left[w_j \cdot |g_j|^2\right]$, it equals $\mathop{\mathbb{E}}_{x_j \sim D_j, y_j \sim Y(x_j)}\left[\alpha_j \cdot \frac{D(x_j)}{D_j(x_j)} \big|y_j - f(x_j)\big|^2\right] = \alpha_j \cdot \mathop{\mathbb{E}}_{x_j \sim D, y_j \sim Y(x_j)}\big[\big|y_j - f(x_j)\big|^2\big]$.

For $\sup_{x_j}\left\{w_j \sum_{i=1}^{d} |v_i(x_j)|^2\right\}$, we bound it as

$$\begin{aligned}\sup_{x_j}\left\{w_j \sum_{i=1}^{d} |v_i(x_j)|^2\right\} &= \sup_{x_j}\left\{\alpha_j \cdot \frac{D(x_j)}{D_j(x_j)} \sum_{i=1}^{d} |v_i(x_j)|^2\right\} \\ &= \alpha_j \sup_{x_j}\left\{\frac{D(x_j)}{D_j(x_j)} \cdot \sup_{h \in \mathcal{F}}\big\{\frac{|h(x_j)|^2}{\|h\|_D^2}\big\}\right\} = \alpha_j \cdot K_{D_j}.\end{aligned}$$

We use the fact $\sup_{h \in \mathcal{F}}\big\{\frac{|h(x_j)|^2}{\|h\|_D^2}\big\} = \sup_{(a_1,\ldots,a_d)}\big\{\frac{|\sum_{i=1}^{d} a_i v_i(x_j)|^2}{\sum_{i=1}^{d} |a_i|^2}\big\} = \frac{(\sum_{i=1}^{d} |a_i|^2)(\sum_{i=1}^{d} |v_i(x_j)|^2)}{\sum_{i=1}^{d} |a_i|^2}$ by the Cauchy-Schwartz inequality. From all discussion above, we have

$$\mathbb{E}[\|A^* \cdot \vec{g}_w\|_2^2] \le \sum_j \left(\alpha_j K_{D_j} \cdot \alpha_j \mathop{\mathbb{E}}_{(x,y) \sim (D,Y)}[|y - f(x)|^2]\right) \le (\sum_j \alpha_j) \max_j \left\{\alpha_j K_{D_j}\right\} \mathop{\mathbb{E}}_{(x,y) \sim (D,Y)}[|y - f(x)|^2].$$

□

We combine all discussion above to prove Theorem 6.0.3.

Proof of Theorem 6.0.3. The first property of P indicates $\lambda(A^* \cdot A) \in [1 - 1/4, 1 + 1/4]$. On the other hand, $\mathbb{E}[\|A^* \cdot (\vec{y}_w - \vec{f}_w)\|_2^2] \leq \epsilon/C \cdot \mathop{\mathbb{E}}_{(x,y)\sim(D,Y)}[|y - f(x)|^2]$ from Lemma 6.2.4. Conditioned on the first property, we have $\mathbb{E}\left[\|(A^* \cdot A)^{-1} \cdot A^* \cdot (\vec{y}_w - \vec{f}_w)\|_2^2\right] \leq 2\epsilon/C \cdot \mathop{\mathbb{E}}_{(x,y)\sim(D,Y)}[|y - f(x)|^2]$.

By choosing a large constant C, with probability $1 - \frac{1}{200}$,

$$\|\widetilde{f} - f\|_D^2 = \|(A^* \cdot A)^{-1} \cdot A^* \cdot (\vec{y}_w - \vec{f}_w)\|_2^2 \leq \lambda_{\max}\big((A^* \cdot A)^{-1}\big) \cdot \|A^* \cdot (\vec{y}_w - \vec{f}_w)\|_2^2 \leq \epsilon \cdot \mathop{\mathbb{E}}_{(x,y)\sim(D,Y)}[|y - f(x)|^2]$$

from the Markov inequality. □

6.2.2 Proof of Corollary 6.2.2

For the first part, let $(D, Y) = \big(D, f(x) + g(x)\big)$ be our joint distribution of (x, y). Because the expectation $\mathbb{E}[g(x)] = 0$ for every $x \in G$, $\arg\min_{v \in V} \mathop{\mathbb{E}}_{(x,y)\sim(D,Y)}[|y - v(x)|^2] = f$. From Theorem 6.0.3, for $\alpha(\widetilde{f}) = (A^* \cdot A)^{-1} \cdot A^* \cdot \vec{y}_w$ and $m = O(d/\varepsilon)$,

$$\|\widetilde{f} - f\|_D^2 = \|\alpha(\widetilde{f}) - \alpha(f)\|_2^2 \leq \varepsilon \cdot \mathop{\mathbb{E}}_{(x,y)\sim(D,Y)}[|y - f(x)|^2] = \varepsilon \cdot \mathbb{E}[\|g\|_D^2], \text{ with probability } 0.99.$$

For the second part, let $g^{\|}$ be the projection of $g(x)$ to $\mathcal{F}$ and $g^{\perp} = g - g^{\|}$ be the orthogonal part to $\mathcal{F}$. Let $\alpha(g^{\|})$ denote the coefficients of $g^{\|}$ in the fixed orthonormal basis $(v_1, \ldots, v_d)$ so that $\|\alpha(g^{\|})\|_2 = \|g^{\|}\|_D$. We decompose $\vec{y}_w = \vec{f}_w + \vec{g}_w = \vec{f}_w + \vec{g^{\|}}_w + \vec{g^{\perp}}_w$. Therefore

$$\alpha(\widetilde{f}) = (A^* A)^{-1} \cdot A^* \cdot (\vec{f}_w + \vec{g^{\|}}_w + \vec{g^{\perp}}_w) = \alpha(f) + \alpha(g^{\|}) + (AA^*)^{-1} A^* \cdot \vec{g^{\perp}}_w.$$

The distance $\|\widetilde{f} - f\|_D = \|\alpha(\widetilde{f}) - \alpha(f)\|_2$ equals

$$\|(A^*A)^{-1}\cdot A^*\cdot\vec{y}_w - \alpha(f)\|_2 = \|\alpha(f)+\alpha(g^{\|})+(A^*A)^{-1}\cdot A^*\cdot\vec{g^{\perp}}_w - \alpha(f)\|_2 = \|\alpha(g^{\|})+(A^*A)^{-1}\cdot A^*\cdot\vec{g^{\perp}}_w\|_2.$$

From Theorem 6.0.3, with probability 0.99, $\|(A^*A)^{-1}\cdot A^*\cdot\vec{g^{\perp}}_w\|_2 \leq \sqrt{\varepsilon}\cdot\|g^{\perp}\|_D$. Thus

$$\begin{aligned}\|(A^*A)^{-1}\cdot A^*\cdot\vec{y}_w - \alpha(f)\|_2 &= \|\alpha(g^{\|}) + (A^*A)^{-1}\cdot A^*\cdot\vec{g^{\perp}}_w\|_2 \\ &\leq \|g^{\|}\|_D + \sqrt{\varepsilon}\cdot\|g^{\perp}\|_D.\end{aligned}$$

Let $1-\beta$ denote $\|g^{\|}\|_D/\|g\|_D$ such that $\|g^{\perp}\|_D/\|g\|_D = \sqrt{2\beta-\beta^2}$. We rewrite it as

$$\left(1-\beta+\sqrt{\varepsilon}\cdot\sqrt{2\beta-\beta^2}\right)\|g\|_D \leq (1-\beta+\sqrt{\varepsilon}\cdot\sqrt{2\beta})\|g\|_D \leq \left(1-(\sqrt{\beta}-\sqrt{\frac{\varepsilon}{2}})^2+\frac{\varepsilon}{2}\right)\|g\|_D.$$

From all discussion above, $\|\widetilde{f} - f\|_D = \|\alpha(\widetilde{f}) - \alpha(f)\|_2 = \|(A^*A)^{-1}\cdot A^*\cdot\vec{y}_w - \alpha(f)\|_2 \leq (1+\varepsilon)\|g\|_D$.

6.3 Performance of i.i.d. Distributions

Given the linear family $\mathcal{F}$ of dimension d and the measure of distance D, we provide a distribution $D_{\mathcal{F}}$ with a condition number $K_{D_{\mathcal{F}}} = d$.

Lemma 6.3.1. *Given any linear family $\mathcal{F}$ of dimension d and any distribution D, there always exists an explicit distribution $D_{\mathcal{F}}$ such that the condition number*

$$K_{D_{\mathcal{F}}} = \sup_x\left\{\sup_{h\in\mathcal{F}}\left\{\frac{D(x)}{D_{\mathcal{F}}(x)}\cdot\frac{|h(x)|^2}{\|h\|_D^2}\right\}\right\} = d.$$

Next, for generality, we bound the number of i.i.d. random samples from an arbitrary distribution D' to fulfill the requirements of *well-balanced sampling procedures* in Definition 6.0.2.

Lemma 6.3.2. *There exists a universal constant C_1 such that given any distribution D' with the same support of D and any $\epsilon > 0$, the random sampling procedure with $m = C_1(K_{D'} \log d + \frac{K_{D'}}{\varepsilon})$ i.i.d. random samples from D' and coefficients $\alpha_1 = \cdots = \alpha_m = 1/m$ is an ε-well-balanced sampling procedure.*

By Theorem 6.0.3, we state the following result, which will be used in active learning. For $G = \mathrm{supp}(D)$ and any $x \in G$, let $Y(x)$ denote the conditional distribution $(Y|D = x)$ and $(D', Y(D'))$ denote the distribution that first generates $x \sim D'$ then generates $y \sim Y(x)$.

Theorem 6.3.3. *Consider any dimension d linear space $\mathcal{F}$ of functions from a domain G to $\mathbb{C}$. Let (D, Y) be a joint distribution over $G \times \mathbb{C}$, and $f = \arg\min_{h \in \mathcal{F}} \mathop{\mathbb{E}}_{(x,y)\sim(D,Y)} [|y - h(x)|^2]$.*

Let D' be any distribution on G and $K_{D'} = \sup_x \left\{ \sup_{h \in \mathcal{F}} \left\{ \frac{D(x)}{D'(x)} \cdot \frac{|h(x)|^2}{\|h\|_D^2} \right\} \right\}$. The weighted ERM $\widetilde{f}$ resulting from $m = O(K_{D'} \log d + \frac{K_{D'}}{\varepsilon})$ random queries of $(D', Y(D'))$ with weights $w_i = \frac{D(x_i)}{m \cdot D'(x_i)}$ for each $i \in [m]$ satisfies

$$\|\tilde{f} - f\|_D^2 = \mathop{\mathbb{E}}_{x \sim D} \left[|\tilde{f}(x) - f(x)|^2 \right] \leq \varepsilon \cdot \mathop{\mathbb{E}}_{(x,y)\sim(D,Y)} \left[|y - f(x)|^2 \right] \text{ with probability } \geq 0.99.$$

We show the proof of Lemma 6.3.2 in Section 6.3.1.

6.3.1 Proof of Lemma 6.3.2

We use the matrix Chernoff theorem to prove the first property in Definition 6.0.2. We still use A to denote the $m \times d$ matrix $A(i, j) = \sqrt{w_i} \cdot v_j(x_i)$.

Lemma 6.3.4. *Let D' be an arbitrary distribution over G and*

$$K_{D'} = \sup_{h \in \mathcal{F}: h \neq 0} \sup_{x \in G} \frac{|h^{(D')}(x)|^2}{\|h\|_D^2}. \tag{6.2}$$

There exists an absolute constant C such that for any $n \in \mathbb{N}^+$, linear family $\mathcal{F}$ of dimension d, $\varepsilon \in (0, 1)$ and $\delta \in (0, 1)$, when $S = (x_1, \ldots, x_m)$ are independently from the distribution

D' with $m \geq \frac{C}{\varepsilon^2} \cdot K_{D'} \log \frac{d}{\delta}$ and $w_j = \frac{D(x_j)}{m \cdot D'(x_j)}$ for each $j \in [m]$, the $m \times d$ matrix $A(i,j) = \sqrt{w_i} \cdot v_j(x_i)$ satisfies

$$\|A^*A - I\| \leq \varepsilon \text{ with probability at least } 1 - \delta.$$

Proof of Lemma 6.3.4. Let $v_1, \ldots, v_d$ be the orthonormal basis of $\mathcal{F}$ in the definition of matrix A. For any $h \in \mathcal{F}$, let $\alpha(h) = (\alpha_1, \ldots, \alpha_d)$ denote the coefficients of h under $v_1, \ldots, v_d$ such that $\|h\|_D^2 = \|\alpha(h)\|_2^2$. At the same time, for any fixed x, $\sup_{h \in \mathcal{F}} \frac{|h^{(D')}(x)|^2}{\|h\|_D^2} = \sup_{\alpha(h)} \frac{|\sum_{i=1}^d \alpha(h)_i \cdot v_i^{(D')}(x)|^2}{\|\alpha(h)\|_2^2} = \sum_{i \in [d]} |v_i^{(D')}(x)|^2$ by the tightness of the Cauchy Schwartz inequality. Thus

$$K_{D'} \stackrel{def}{=} \sup_{x \in G} \Big\{ \sup_{h \in \mathcal{F}: h \neq 0} \frac{|h^{(D')}(x)|^2}{\|h\|_D^2} \Big\} \quad \text{indicates} \quad \sup_{x \in G} \sum_{i \in [d]} |v_i^{(D')}(x)|^2 \leq K_{D'}. \tag{6.3}$$

For each point x_j in S with weight $w_j = \frac{D(x_j)}{m \cdot D'(x_j)}$, let A_j denote the jth row of the matrix A. It is a vector in $\mathbb{C}^d$ defined by $A_j(i) = A(j,i) = \sqrt{w_j} \cdot v_i(x_j) = \frac{v_i^{(D')}(x_j)}{\sqrt{m}}$. So $A^*A = \sum_{j=1}^m A_j^* \cdot A_j$.

For $A_j^* \cdot A_j$, it is always $\succeq 0$. Notice that the only non-zero eigenvalue of $A_j^* \cdot A_j$ is

$$\lambda(A_j^* \cdot A_j) = A_j \cdot A_j^* = \frac{1}{m} \left(\sum_{i \in [d]} |v_i^{(D')}(x_j)|^2 \right) \leq \frac{K_{D'}}{m}$$

from (6.3).

At the same time, $\sum_{j=1}^m \mathbb{E}[A_j^* \cdot A_j]$ equals the identity matrix of size $d \times d$ because the expectation of the entry (i,i') in $A_j^* \cdot A_j$ is

$$\begin{aligned} \mathop{\mathbb{E}}_{x_j \sim D'} [\overline{A(j,i)} \cdot A(j,i')] &= \mathop{\mathbb{E}}_{x_j \sim D'} [\frac{\overline{v_i^{(D')}(x_j)} \cdot v_{i'}^{(D')}(x_j)}{m}] \\ &= \mathop{\mathbb{E}}_{x_j \sim D'} [\frac{D(x) \cdot \overline{v_i(x_j)} \cdot v_{i'}(x_j)}{m \cdot D'(x_j)}] = \mathop{\mathbb{E}}_{x_j \sim D} [\frac{\overline{v_i(x_j)} \cdot v_{i'}(x_j)}{m}] = 1_{\vec{i} = \vec{i'}}/m. \end{aligned}$$

Now we apply Theorem 2.2.3 on $A^*A = \sum_{j=1}^{m}(A_j^* \cdot A_j)$:

$$\begin{aligned}\Pr\left[\lambda(A^*A) \notin [1-\varepsilon, 1+\varepsilon]\right] &\leq d\left(\frac{e^{-\varepsilon}}{(1-\varepsilon)^{1-\varepsilon}}\right)^{1/\frac{K_{D'}}{m}} + d\left(\frac{e^{-\varepsilon}}{(1+\varepsilon)^{1+\varepsilon}}\right)^{1/\frac{K_{D'}}{m}} \\ &\leq 2d \cdot e^{-\frac{\varepsilon^2 \cdot \frac{m}{K_{D'}}}{3}} \leq \delta \qquad \text{given } m \geq \frac{6K_{D'}\log\frac{d}{\delta}}{\varepsilon^2}.\end{aligned}$$

□

Then we finish the proof of Lemma 6.3.2.

Proof of Lemma 6.3.2. Because the coefficient $\alpha_i = 1/m = \Omega(K_{D'}/\varepsilon)$ and $\sum_i \alpha_i = 1$, this indicates the second property of *well-balanced sampling procedures*.

Since $m = \Theta(K_{D'}\log d)$, by Lemma 6.3.4, we know all eigenvalues of $A^* \cdot A$ are in $[1-1/4, 1+1/4]$ with probability $1-10^{-3}$. By Lemma 6.2.3, this indicates the first property of *well-balanced sampling procedures*. □

6.4 A Linear-Sample Algorithm for Known D

We provide a well-balanced sampling procedure with a linear number of random samples in this section. The procedure requires knowing the underlying distribution D, which makes it directly useful in the query setting or the "fixed design" active learning setting, where D can be set to the empirical distribution D_0.

Lemma 6.4.1. *Given any dimension d linear space $\mathcal{F}$, any distribution D over the domain of $\mathcal{F}$, and any $\varepsilon > 0$, there exists an efficient* ε-well-balanced sampling procedure *that terminates in $O(d/\varepsilon)$ rounds with probability $1 - \frac{1}{200}$.*

From Corollary 6.2.2, we obtain the following theorem for the ERM $\widetilde{f}$ resulting from the *well-balanced sampling procedure* in Lemma 6.4.1.

Theorem 6.4.2. *Consider any dimension d linear space $\mathcal{F}$ of functions from a domain G to $\mathbb{C}$ and distribution D over G. Let $y(x) = f(x) + g(x)$ be our observed function, where $f \in \mathcal{F}$ and g denotes a noise function. For any $\varepsilon > 0$, there exists an efficient algorithm that observes $y(x)$ at $m = O(\frac{d}{\varepsilon})$ points and outputs $\widetilde{f}$ such that with probability 0.99,*

1. *$\|\tilde{f} - f\|_D^2 \leq \varepsilon \cdot \mathbb{E}_g[\|g\|_D^2]$, when $g(x)$ is a random function from G to $\mathbb{C}$ where each $g(x)$ is an independent random variable with $\mathbb{E}_g[g(x)] = 0$.*
2. *$\|\tilde{f} - f\|_D \leq (1+\varepsilon) \cdot \|g\|_D$ for any other noise function g.*

We show how to extract the coefficients $\alpha_1, \cdots, \alpha_m$ from the randomized BSS algorithm by Lee and Sun [LS15] in Algorithm 4. Given ϵ, the linear family $\mathcal{F}$, and the distribution D, we fix $\gamma = \sqrt{\epsilon}/C_0$ for a constant C_0 and $v_1, \ldots, v_d$ to be an orthonormal basis of $\mathcal{F}$ in this section. For convenience, we use $v(x)$ to denote the vector $\big(v_1(x), \ldots, v_d(x)\big)$.

In the rest of this section, we prove Lemma 6.4.1 in Section 6.4.1.

6.4.1 Proof of Lemma 6.4.1

We state a few properties of randomized BSS [BSS12, LS15] that will be used in this proof. The first property is that matrices $B_1, \ldots, B_m$ in Procedure RandomizedBSS always have bounded eigenvalues.

Lemma 6.4.3. *[BSS12, LS15] For any $j \in [m]$, $\lambda(B_j) \in (l_j, u_j)$.*

Lemma 3.6 and 3.7 of [LS15] shows that with high probability, the while loop in Procedure RandomizedSamplingBSS finishes within $O(\frac{d}{\gamma^2})$ iterations and guarantees the last matrix B_m is well-conditioned, i.e., $\frac{\lambda_{\max}(B_m)}{\lambda_{\min}(B_m)} \leq \frac{u_m}{l_m} \leq 1 + O(\gamma)$.

Lemma 6.4.4. *[LS15] There exists a constant C such that with probability at least $1 - \frac{1}{200}$, Procedure* RandomizedSamplingBSS *takes at most $m = C \cdot d/\gamma^2$ random points $x_1, \ldots, x_m$ and guarantees that $\frac{u_m}{l_m} \leq 1 + 8\gamma$.*

Algorithm 4 A well-balanced sampling procedure based on Randomized BSS

procedure RandomizedSamplingBSS($\mathcal{F}, D, \epsilon$)
  Find an orthonormal basis $v_1, \ldots, v_d$ of $\mathcal{F}$ under D;
  Set $\gamma = \sqrt{\epsilon}/C_0$ and mid $= \frac{4d/\gamma}{1/(1-\gamma)-1/(1+\gamma)}$;
  $j = 0; B_0 = 0$;
  $l_0 = -2d/\gamma; u_0 = 2d/\gamma$;
  while $u_{j+1} - l_{j+1} < 8d/\gamma$ **do**;
    $\Phi_j = \mathrm{Tr}(u_j I - B_j)^{-1} + \mathrm{Tr}(B_j - l_j I)^{-1}$; ▷ The potential function at iteration j.
    Set the coefficient $\alpha_j = \frac{\gamma}{\Phi_j} \cdot \frac{1}{\text{mid}}$;
    Set the distribution $D_j(x) = D(x) \cdot \Big(v(x)^\top (u_j I - B_j)^{-1} v(x) + v(x)^\top (B_j - l_j I)^{-1} v(x) \Big) / \Phi_j$ for $v(x) = \big(v_1(x), \ldots, v_d(x)\big)$;
    Sample $x_j \sim D_j$ and set a scale $s_j = \frac{\gamma}{\Phi_j} \cdot \frac{D(x)}{D_j(x)}$;
    $B_{j+1} = B_j + s_j \cdot v(x_j) v(x_j)^\top$;
    $u_{j+1} = u_j + \frac{\gamma}{\Phi_j(1-\gamma)}$; $l_{j+1} = l_j + \frac{\gamma}{\Phi_j(1+\gamma)}$;
    $j = j + 1$;
  end while
  $m = j$;
  Assign the weight $w_j = s_j / \text{mid}$ for each x_j;
end procedure

We first show that $(A^* \cdot A)$ is well-conditioned from the definition of A. We prove that our choice of mid is very close to $\sum_{j=1}^{m} \frac{\gamma}{\phi_j} = \frac{u_m + l_m}{\frac{1}{1-\gamma} + \frac{1}{1+\gamma}} \approx \frac{u_m + l_m}{2}$.

Claim 6.4.5. *After exiting the while loop in Procedure* RandomizedBSS*, we always have*

1. $u_m - l_m \leq 9d/\gamma$.
2. $(1 - \frac{0.5\gamma^2}{d}) \cdot \sum_{j=1}^{m} \frac{\gamma}{\phi_j} \leq \text{mid} \leq \sum_{j=1}^{m} \frac{\gamma}{\phi_j}$.

Proof. Let us first bound the last term $\frac{\gamma}{\phi_m}$ in the while loop. Since $u_{m-1} - l_{m-1} < 8d/\gamma$, $\phi_m \geq 2d \cdot \frac{1}{4d/\gamma} \geq \frac{\gamma}{2}$, which indicates the last term $\frac{\gamma}{\phi_m} \leq 2$. Thus

$$u_m - l_m \leq 8d/\gamma + 2(\frac{1}{1-\gamma} - \frac{1}{1+\gamma}) \leq 8d/\gamma + 5\gamma.$$

From our choice $\text{mid} = \frac{4d/\gamma}{1/(1-\gamma) - 1/(1+\gamma)} = 2d(1-\gamma^2)/\gamma^2$ and the condition of the while loop $u_m - l_m = \sum_{j=1}^{m} (\gamma/\phi_j) \cdot (\frac{1}{1-\gamma} - \frac{1}{1+\gamma}) + 4d/\gamma \geq 8d/\gamma$, we know

$$\sum_{j=1}^{m} \frac{\gamma}{\phi_j} \geq \text{mid} = 2d(1-\gamma^2)/\gamma^2.$$

On the other hand, since $u_{m-1} - l_{m-1} < 8d/\gamma$ is in the while loop, $\sum_{j=1}^{m-1} \frac{\gamma}{\phi_j} < \text{mid}$. Hence

$$\text{mid} > \sum_{j=1}^{m-1} \frac{\gamma}{\phi_j} \geq \sum_{j=1}^{m} \frac{\gamma}{\phi_j} - 2 \geq (1 - 0.5\gamma^2/d) \cdot (\sum_{j=1}^{m} \frac{\gamma}{\phi_j}).$$

□

Lemma 6.4.6. *Given* $\frac{u_m}{l_m} \leq 1 + 8\gamma$, $\lambda(A^* \cdot A) \in (1 - 5\gamma, 1 + 5\gamma)$.

Proof. For $B_m = \sum_{j=1}^{m} s_j v(x_j) v(x_j)^\top$, $\lambda(B_m) \in (l_m, u_m)$ from Lemma 6.4.3. At the same time, given $w_j = s_j / \text{mid}$,

$$(A^* A) = \sum_{j=1}^{m} w_j v(x_j) v(x_j)^\top = \frac{1}{\text{mid}} \cdot \sum_{j=1}^{m} s_j v(x_j) v(x_j)^\top = \frac{B_m}{\text{mid}}.$$

Since $\text{mid} \in [1-\frac{3\gamma^2}{d}, 1] \cdot (\sum_{j=1}^{m} \frac{\gamma}{\phi_j}) = [1-\frac{3\gamma^2}{d}, 1] \cdot (\frac{u_m+l_m}{\frac{1}{1-\gamma}+\frac{1}{1+\gamma}}) \subseteq [1-2\gamma^2, 1-\gamma^2] \cdot (\frac{u_m+l_m}{2})$ from Claim 6.4.5, $\lambda(A^* \cdot A) = \lambda(B_m)/\,\text{mid} \in (l_m/\,\text{mid}, u_m/\,\text{mid}) \subset (1-5\gamma, 1+5\gamma)$ given $\frac{u_m}{l_m} \leq 1+8\gamma$ in Lemma 6.4.4. □

We finish the proof of Lemma 6.4.1 by combining all discussion above.

Proof of Lemma 6.4.1. From Lemma 6.4.4 and Lemma 6.4.6, $m = O(d/\gamma^2)$ and $\lambda(A^*A) \in [1-1/4, 1+1/4]$ with probability 0.995.

For $\alpha_i = \frac{\gamma}{\Phi_i} \cdot \frac{1}{\text{mid}}$, we bound $\sum_{i=1}^{m} \frac{\gamma}{\Phi_i} \cdot \frac{1}{\text{mid}}$ by 1.25 from the second property of Claim 6.4.5.

Then we bound $\alpha_j \cdot K_{D_j}$. We notice that $\sup_{h\in\mathcal{F}} \frac{|h(x)|^2}{\|h\|_D^2} = \sum_{i\in[d]} |v_i(x)|^2$ for every $x \in G$ because $\sup_{h\in\mathcal{F}} \frac{|h(x)|^2}{\|h\|_D^2} = \sup_{\alpha(h)} \frac{\left|\sum_i \alpha(h)_i \cdot v_i(x)\right|^2}{\|\alpha(h)\|_2^2} = \sum_i |v_i(x)|^2$ by the Cauchy-Schwartz inequality. This simplifies K_{D_j} to $\sup_x\{\frac{D(x)}{D_j(x)} \cdot \sum_{i=1}^{d} |u_i(x)|^2\}$ and bounds $\alpha_j \cdot K_{D_j}$ by

$$
\begin{aligned}
&\frac{\gamma}{\Phi_j \cdot \text{mid}} \cdot \sup_x \left\{ \frac{D(x)}{D_j(x)} \cdot \sum_{i=1}^{d} |v_i(x)|^2 \right\} \\
=&\frac{\gamma}{\text{mid}} \cdot \sup_x \left\{ \frac{\sum_{i=1}^{d} |v_i(x)|^2}{v(x_j)^\top (u_j I - B_j)^{-1} v(x_j) + v(x_j)^\top (B_j - l_j I)^{-1} v(x_j)} \right\} \\
\leq&\frac{\gamma}{\text{mid}} \cdot \sup_x \left\{ \frac{\sum_{i=1}^{d} |v_i(x)|^2}{\lambda_{\min}\big((u_j I - B_j)^{-1}\big) \cdot \|v(x_j)\|_2^2 + \lambda_{\min}\big((B_j - l_j I)^{-1}\big) \cdot \|v(x_j)\|_2^2} \right\} \\
\leq&\frac{\gamma}{\text{mid}} \cdot \frac{1}{1/(u_j - l_j) + 1/(u_j - l_j)} \\
=&\frac{\gamma}{\text{mid}} \cdot \frac{u_j - l_j}{2} \qquad \text{(apply the first property of Claim 6.4.5)} \\
\leq&\frac{4.5 \cdot d}{\text{mid}} \leq 3\gamma^2 = 3\epsilon/C_0^2.
\end{aligned}
$$

By choosing C_0 large enough, this satisfies the second property of *well-balanced sampling procedures*. At the same time, by Lemma 6.2.3, Algorithm 4 also satisfies the first property of *well-balanced sampling procedures*. □

6.5 Active Learning

In this section, we investigate the case where we do not know the distribution D of x and only receive random samples from D. We finish the proof of Theorem 6.0.4 that bounds the number of unlabeled samples by the condition number of D and the number of labeled samples by $dim(\mathcal{F})$ to find the truth through D. In the end of this section, we state a corollary for specific kinds of noise.

Theorem 6.0.4. *Consider any dimension d linear space $\mathcal{F}$ of functions from a domain G to $\mathbb{C}$. Let (D, Y) be a joint distribution over $G \times \mathbb{C}$ and $f = \arg\min_{h \in \mathcal{F}} \mathop{\mathbb{E}}_{(x,y)\sim(D,Y)}[|y - h(x)|^2]$.*

Let $K = \sup_{h \in \mathcal{F}: h \neq 0} \frac{\sup_{x \in G} |h(x)|^2}{\|h\|_D^2}$. For any $\varepsilon > 0$, there exists an efficient algorithm that takes $O(K \log d + \frac{K}{\varepsilon})$ unlabeled samples from D and requests $O(\frac{d}{\varepsilon})$ labels to output $\widetilde{f}$ satisfying

$$\mathop{\mathbb{E}}_{x\sim D}[|\tilde{f}(x) - f(x)|^2] \leq \varepsilon \cdot \mathop{\mathbb{E}}_{(x,y)\sim(D,Y)}[|y - f(x)|^2] \text{ with probability } \geq 0.99.$$

For generality, we bound the number of labels using any *well-balanced sampling procedure*, such that Theorem 6.0.4 follows from this lemma with the linear sample procedure in Lemma 6.4.1.

Lemma 6.5.1. *Consider any dimension d linear space $\mathcal{F}$ of functions from a domain G to $\mathbb{C}$. Let (D, Y) be a joint distribution over $G \times \mathbb{C}$ and $f = \arg\min_{h \in \mathcal{F}} \mathop{\mathbb{E}}_{(x,y)\sim(D,Y)}[|y - h(x)|^2]$.*

Let $K = \sup_{h \in \mathcal{F}: h \neq 0} \frac{\sup_{x \in G} |h(x)|^2}{\|h\|_D^2}$ and P be a well-balanced sampling procedure *terminating in $m_p(\varepsilon)$ rounds with probability $1 - 10^{-3}$ for any linear family $\mathcal{F}$, measurement D, and ε. For any $\varepsilon > 0$, Algorithm 5 takes $O(K \log d + \frac{K}{\varepsilon})$ unlabeled samples from D and requests at most $m_p(\varepsilon/8)$ labels to output $\widetilde{f}$ satisfying*

$$\mathop{\mathbb{E}}_{x\sim D}[|\tilde{f}(x) - f(x)|^2] \leq \varepsilon \cdot \mathop{\mathbb{E}}_{(x,y)\sim(D,Y)}[|y - f(x)|^2] \text{ with probability } \geq 1 - \frac{1}{200}.$$

Algorithm 5 first takes $m_0 = O(K \log d + K/\varepsilon)$ unlabeled samples and defines a distribution D_0 to be the uniform distribution on these m_0 samples. Then it uses D_0 to simulate D in P, i.e., it outputs the ERM resulting from the well-balanced sampling procedure P with the linear family $\mathcal{F}$, the measurement D_0, and $\frac{\epsilon}{8}$.

Algorithm 5 Regression over an unknown distribution D

1: **procedure** RegressionUnknownDistribution($\varepsilon, \mathcal{F}, D, P$)
2: Set C to be a large constant and $m_0 = C \cdot (K \log d + K/\varepsilon)$.
3: Take m_0 unlabeled samples $x_1, \ldots, x_{m_0}$ from D.
4: Let D_0 be the uniform distribution over $(x_1, \ldots, x_{m_0})$.
5: Output the ERM $\tilde{f}$ resulting from P with parameters $\mathcal{F}, D_0, \epsilon/8$.
6: **end procedure**

Proof. We still use $\|f\|_{D'}$ to denote $\sqrt{\mathbb{E}_{x \sim D'}[|f(x)|^2]}$ and D_1 to denote the weighted distribution generated by Procedure P given $\mathcal{F}, D_0, \varepsilon$. By Lemma 6.3.2 with D and the property of P, with probability at least $1 - 2 \cdot 10^{-3}$,

$$\|h\|_{D_0}^2 = (1 \pm 1/4) \cdot \|h\|_D^2 \text{ and } \|h\|_{D_1}^2 = (1 \pm 1/4) \cdot \|h\|_{D_0}^2 \text{ for every } h \in \mathcal{F}. \tag{6.4}$$

We assume (6.4) holds in the rest of this proof.

Let y_i denote a random label of x_i from $Y(x_i)$ for each $i \in [m_0]$ including the unlabeled samples in the algorithm and the labeled samples in Step 5 of Algorithm 5. Let f' be the weighted ERM of $(x_1, \cdots, x_m)$ and $(y_1, \cdots, y_m)$ over D_0, i.e.,

$$f' = \arg\min_{h \in \mathcal{F}} \mathop{\mathbb{E}}_{x_i \sim D_0, y_i \sim Y(x_i)} \left[|y_i - h(x_i)|^2 \right]. \tag{6.5}$$

Given Property (6.4) and Lemma 6.3.2, $\mathop{\mathbb{E}}_{(x_1,y_1),\ldots,(x_{m_0},y_{m_0})}[\|f' - f\|_D^2] \leq \frac{2K}{m_0} \cdot \mathop{\mathbb{E}}_{(x,y) \sim (D,Y)}[|y - f(x)|^2]$ from the proof of Theorem 6.0.3. Let the constant in m_0 be large enough such that from the Markov inequality, with probability $1 - 10^{-3}$,

$$\|f' - f\|_D^2 \leq \frac{\varepsilon}{8} \cdot \mathop{\mathbb{E}}_{(x,y) \sim (D,Y)}[|y - f(x)|^2].$$

In the rest of this proof, we plan to show that the weighted ERM $\widetilde{f}$ resulting from P with measurement D_0 guarantees $\|\widetilde{f}-f'\|_{D_0}^2 \lesssim \mathop{\mathbb{E}}_{(x,y)\sim(D,Y)}[|y-f(x)|^2]$ with high probability. Given Property (6.4) and the guarantee of Procedure P, we have $\mathop{\mathbb{E}}_{P}[\|\widetilde{f}-f'\|_{D_0}^2] \le \frac{2\epsilon}{C}\cdot \mathop{\mathbb{E}}_{x\sim D_0}[|y_i-f'(x_i)|^2]$ from the proof of Theorem 6.0.3. Next we bound the right hand side $\mathop{\mathbb{E}}_{x_i\sim D_0}[|y_i-f'(x_i)|^2]$ by $\mathop{\mathbb{E}}_{(x,y)\sim(D,Y)}[|y-f(x)|^2]$ over the randomness of $(x_1,y_1),\dots,(x_{m_0},y_{m_0})$:

$$
\begin{aligned}
&\mathop{\mathbb{E}}_{(x_1,y_1),\dots,(x_{m_0},y_{m_0})}\left[\mathop{\mathbb{E}}_{x_i\sim D_0}\left[|y_i-f'(x_i)|^2\right]\right]\\
\le&\mathop{\mathbb{E}}_{(x_1,y_1),\dots,(x_{m_0},y_{m_0})}\left[2\mathop{\mathbb{E}}_{x_i\sim D_0}\left[|y_i-f(x_i)|^2\right]+2\|f-f'\|_{D_0}^2\right]\\
\le&2\mathop{\mathbb{E}}_{(x,y)\sim(D,Y)}\left[|y-f(x)|^2\right]+3\mathop{\mathbb{E}}_{(x_1,y_1),\dots,(x_{m_0},y_{m_0})}\left[\|f-f'\|_D^2\right] \quad \text{from (6.4)}
\end{aligned}
$$

Hence $\mathop{\mathbb{E}}_{(x_1,y_1),\dots,(x_{m_0},y_{m_0})}\left[\mathop{\mathbb{E}}_{P}[\|\widetilde{f}-f'\|_{D_0}^2]\right] \le \frac{2\epsilon}{C}\cdot\left(2+3\cdot\frac{2K}{m_0}\right)\cdot\mathop{\mathbb{E}}_{(x,y)\sim(D,Y)}[|y-f(x)|^2]$.

Since C is a large constant, from the Markov inequality, with probability $1-2\cdot10^{-3}$ over $(x_1,y_1),\dots,(x_{m_0},y_{m_0})$ and P,

$$\|\widetilde{f}-f'\|_{D_0}^2 \le \frac{\varepsilon}{4}\cdot\mathop{\mathbb{E}}_{(x,y)\sim(D,Y)}\left[|y-f(x)|^2\right].$$

From all discussion above, we have

$$\|\widetilde{f}-f\|_D^2 \le 2\|\widetilde{f}-f'\|_D^2+2\|f'-f\|_D^2 \le 3\|\widetilde{f}-f'\|_{D_0}^2+\frac{\varepsilon}{4}\mathop{\mathbb{E}}_{(x,y)\sim(D,Y)}[|y-f(x)|^2] \le \varepsilon\mathop{\mathbb{E}}_{(x,y)\sim(D,Y)}[|y-f(x)|^2].$$

□

We apply Corollary 6.2.2 to obtain the following result.

Corollary 6.5.2. *Let $\mathcal{F}$ be a family of functions from a domain G to $\mathbb{C}$ with dimension d and D be a distribution over G with bounded condition number $K=\sup_{h\in\mathcal{F}:h\neq0}\frac{\sup_{x\in G}|h(x)|^2}{\|h\|_D^2}$. Let $y(x)=f(x)+g(x)$ be our observation with $f\in\mathcal{F}$.*

For any $\varepsilon>0$, there exists an efficient algorithm that takes $O(K\log d+\frac{K}{\varepsilon})$ unlabeled samples from D and require $O(\frac{d}{\varepsilon})$ labels to output $\widetilde{f}$ such that with probability 0.99,

1. $\|\widetilde{f} - f\|_D^2 \le \varepsilon \cdot \mathbb{E}[\|g\|_D^2]$, *when* g *is a random function where each* $g(x)$ *is an independent random variable with* $\mathbb{E}[g(x)] = 0$.

2. $\|\widetilde{f} - f\|_D \le (1+\varepsilon) \cdot \|g\|_D$ *otherwise.*

6.6 Lower Bounds

We present two lower bounds on the number of samples in this section. We first prove a lower bound for query complexity based on the dimension d. Then we prove a lower bound for the sample complexity based on the condition number of the sampling distribution.

Theorem 6.6.1. *For any* d *and any* $\varepsilon < \frac{1}{10}$, *there exist a distribution* D *and a linear family* $\mathcal{F}$ *of functions with dimension* d *such that for the i.i.d. Gaussian noise* $g(x) = N(0, \frac{1}{\varepsilon})$, *any algorithm which observes* $y(x) = f(x) + g(x)$ *for* $f \in \mathcal{F}$ *with* $\|f\|_D = 1$ *and outputs* $\widetilde{f}$ *satisfying* $\|f - \widetilde{f}\|_D \le 0.1$ *with probability* $\ge \frac{3}{4}$, *needs at least* $m \ge \frac{0.8d}{\varepsilon}$ *queries.*

Notice that this lower bound matches the upper bound in Theorem 6.4.2. In the rest of this section, we focus on the proof of Theorem 6.6.1. Let $\mathcal{F} = \{f : [d] \to \mathbb{R}\}$ and D be the uniform distribution over $[d]$. We first construct a packing set $\mathcal{M}$ of $\mathcal{F}$.

Claim 6.6.2. *There exists a subset* $\mathcal{M} = \{f_1, \ldots, f_n\} \subseteq \mathcal{F}$ *with the following properties:*

1. $\|f_i\|_D = 1$ *for each* $f_i \in \mathcal{M}$.

2. $\|f_i\|_\infty \le 1$ *for each* $f_i \in \mathcal{M}$.

3. $\|f_i - f_j\|_D > 0.2$ *for distinct* f_i, f_j *in* $\mathcal{M}$.

4. $n \ge 2^{0.7d}$.

Proof. We construct $\mathcal{M}$ from $U = \{f : [d] \to \{\pm 1\}\}$ in Procedure ConstructM. Notice that $|U| = 2^d$ before the while loop. At the same time, Procedure ConstructM removes at most $\binom{d}{\leq 0.01d} \leq 2^{0.3d}$ functions every time because $\|g - h\|_D < 0.2$ indicates $\Pr[g(x) \neq h(x)] \leq (0.2)^2/4 = 0.01$. Thus $n \geq 2^d/2^{0.3d} \geq 2^{0.7d}$.

Algorithm 6 Construct $\mathcal{M}$

procedure ConstructM(d)
    Set $n = 0$ and $U = \{f : [d] \to \{\pm 1\}\}$.
    while $U \neq \emptyset$ **do**
        Choose any $h \in U$ and remove all functions $h' \in U$ with $\|h - h'\|_D < 0.2$.
        $n = n + 1$ and $f_n = h$.
    end while
    Return $\mathcal{M} = \{f_1, \ldots, f_n\}$.
end procedure

□

We state the Shannon-Hartley theorem in information theory to finish the proof of Theorem 6.6.1.

Theorem 6.6.3 (The Shannon-Hartley Theorem [Har28, Sha49])**.** *Let S be a real-valued random variable with $\mathbb{E}[S^2] = \tau^2$ and $T \sim N(0, \sigma^2)$. The mutual information $I(S; S + T) \leq \frac{1}{2}\log(1 + \frac{\tau^2}{\sigma^2})$.*

Proof of Theorem 6.6.1. Because of Yao's minimax principle, we assume A is a deterministic algorithm given the i.i.d. Gaussian noise. Let $I(\widetilde{f}; f_j)$ denote the mutual information of a random function $f_j \in \mathcal{M}$ and A's output $\widetilde{f}$ given m observations $(x_1, y_1), \ldots, (x_m, y_m)$ with $y_i = f_j(x_i) + N(0, \frac{1}{\varepsilon})$. When the output $\widetilde{f}$ satisfies $\|\widetilde{f} - f_j\|_D \leq 0.1$, f_j is the closest function to $\widetilde{f}$ in $\mathcal{M}$ from the third property of $\mathcal{M}$. From Fano's inequality [Fan61],

$H(f_j|\tilde{f}) \le H(\frac{1}{4}) + \frac{\log(|\mathcal{M}|-1)}{4}$. This indicates

$$I(f_j; \widetilde{f}) = H(f_j) - H(f_j|\tilde{f}) \ge \log|\mathcal{M}| - 1 - \log(|\mathcal{M}|-1)/4 \ge 0.7\log|\mathcal{M}| \ge 0.4d.$$

At the same time, by the data processing inequality, the algorithm A makes m queries $\big(x_1, \ldots, x_m\big)$ and sees $\big(y_1, \ldots, y_m\big)$, which indicates

$$I(\widetilde{f}; f_j) \le I\bigg(\big(y_1, \ldots, y_m\big); f_j\bigg) = \sum_{i=1}^{m} I\bigg(y_i; f_j(x_i)\big|y_1, \cdots, y_{i-1}\bigg). \tag{6.6}$$

For the query x_i, let $D_{i,j}$ denote the distribution of $f_j \in \mathcal{M}$ in the algorithm A given the first $i-1$ observations $\big(x_1, y_1\big), \ldots, \big(x_{i-1}, y_{i-1}\big)$. We apply Theorem 6.6.3 on $D_{i,j}$ such that it bounds

$$\begin{aligned} I\left(y_i = f_j(x_i) + N(0, \frac{1}{\varepsilon}); f_j(x_i)\big|y_1, \cdots, y_{i-1}\right) \le & \frac{1}{2}\log\left(1 + \frac{\mathop{\mathbb{E}}_{f\sim D_{i,j}}[f(x_i)^2]}{1/\epsilon}\right) \\ \le & \frac{1}{2}\log\left(1 + \frac{\max_{f\in\mathcal{M}}[f(x_i)^2]}{1/\varepsilon}\right) \\ = & \frac{1}{2}\log\big(1+\varepsilon\big) \le \frac{\varepsilon}{2}, \end{aligned}$$

where we apply the second property of $\mathcal{M}$ in the second step to bound $f(x)^2$ for any $f \in \mathcal{M}$. Hence we bound $\sum_{i=1}^{m} I(y_i; f_j|y_1, \cdots, y_{i-1})$ by $m \cdot \frac{\varepsilon}{2}$. This implies

$$0.4d \le m \cdot \frac{\varepsilon}{2} \Rightarrow m \ge \frac{0.8d}{\varepsilon}.$$

□

Next we consider the sample complexity of linear regression.

Theorem 6.6.4. *For any K, d, and $\varepsilon > 0$, there exist a distribution D, a linear family of functions $\mathcal{F}$ with dimension d whose condition number $\sup_{h\in\mathcal{F}:h\neq 0}\left\{\sup_{x\in G}\frac{|h(x)|^2}{\|h\|_D^2}\right\}$ equals K, and a noise function g orthogonal to V such that any algorithm observing $y(x) = f(x) + g(x)$ of $f \in \mathcal{F}$ needs at least $\Omega(K\log d + \frac{K}{\varepsilon})$ samples from D to output $\widetilde{f}$ satisfying $\|\widetilde{f} - f\|_D \leq 0.1\sqrt{\varepsilon}\cdot\|g\|_D$ with probability $\frac{3}{4}$.*

Proof. We fix K to be an integer and set the domain of functions in $\mathcal{F}$ to be $[K]$. We choose D to be the uniform distribution over $[K]$. Let $\mathcal{F}$ denote the family of functions $\big\{f : [K] \to \mathbb{C} | f(d+1) = f(d+2) = \cdots = f(K) = 0\big\}$. Its condition number $\sup_{h\in\mathcal{F}:h\neq 0}\big\{\sup_{x\in G}\frac{|h(x)|^2}{\|h\|_D^2}\big\}$ equals K. $h(x) = 1_{x=1}$ provides the lower bound $\geq K$. At the same time, $\frac{|h(x)|^2}{\|h\|_D^2} = \frac{|h(x)|^2}{\sum_{i=1}^K |h(x)|^2/K} \leq K$ indicates the upper bound $\leq K$.

We first consider the case $K\log d \geq \frac{K}{\varepsilon}$. Let $g = 0$ such that g is orthogonal to V. Notice that $\|\widetilde{f} - f\|_D \leq 0.1\sqrt{\varepsilon}\cdot\|g\|_D$ indicates $\widetilde{f}(x) = f(x)$ for every $x \in [d]$. Hence the algorithm needs to sample $f(x)$ for every $x \in [d]$ when sampling from D: the uniform distribution over $[K]$. From the lower bound of the coupon collector problem, this takes at least $\Omega(K\log d)$ samples from D.

Otherwise, we prove that the algorithm needs $\Omega(K/\varepsilon)$ samples. Without loss of generality, we assume $\mathop{\mathbb{E}}_{x\sim[d]}\big[|f(x)|^2\big] = 1$ for the truth f in y. Let $g(x) = N(0, 1/\varepsilon)$ for each $x \in [d]$. From Theorem 6.6.1, to find $\widetilde{f}$ satisfying $\mathop{\mathbb{E}}_{x\sim[d]}\big[|\widetilde{f}(x) - f(x)|^2\big] \leq 0.1\mathop{\mathbb{E}}_{x\sim[d]}\big[|f(x)|^2\big]$, the algorithm needs at least $\Omega(d/\varepsilon)$ queries of $x \in [d]$. Hence it needs $\Omega(K/\varepsilon)$ random samples from D, the uniform distribution over $[K]$. □

Chapter 7

Existence of Extractors in Simple Hash Families

We study randomness extractors consisting of hash families in this section. Recall that the min-entropy of a random variable X is

$$H_\infty(X) = \min_{x \in \mathrm{supp}(X)} \log_2 \frac{1}{\Pr[X = x]}.$$

For convenience, we provide the definition of extractors and strong extractors here.

Definition 7.0.1. *For any $d \in \mathbb{N}^+$, let U_d denote the uniform distribution over $\{0,1\}^d$. For two random variables W and Z with the same support, let $\|W - Z\|$ denote the statistical (variation) distance*

$$\|W - Z\| = \max_{T \subseteq \mathrm{supp}(W)} \big| \Pr_{w \sim W}[w \in T] - \Pr_{z \sim Z}[z \in T] \big|.$$

A function $\mathsf{Ext} : \{0,1\}^n \times \{0,1\}^t \to \{0,1\}^m$ is a (k,ϵ)-extractor if for every source X with min-entropy k and an independent uniform distribution Y on $\{0,1\}^t$,

$$\|\mathsf{Ext}(X,Y) - U_m\| \leq \epsilon.$$

It is a strong (k,ϵ)-extractor if in addition, it satisfies $\big\|\big(\mathsf{Ext}(X,Y), Y\big) - \big(U_m, Y\big)\big\| \leq \epsilon$.

Given an extractor $\mathsf{Ext} : \{0,1\}^n \times \{0,1\}^t \to \{0,1\}^m$, we sample a few seeds from $\{0,1\}^t$ and consider the new extractor, called a restricted extractor, constituted by these seeds.

Definition 1.1.3. *Given an extractor* $\mathsf{Ext} : \{0,1\}^n \times \{0,1\}^t \to \{0,1\}^m$ *and a sequence of seeds* $(y_1, \cdots, y_D)$ *where each* $y_i \in \{0,1\}^t$*, we define the restricted extractor* $\mathsf{Ext}_{(y_1,\cdots,y_D)}$ *to be* Ext *restricted in the domain* $\{0,1\}^n \times [D]$ *where* $\mathsf{Ext}_{(y_1,\cdots,y_D)}(x,i) = \mathsf{Ext}(x, y_i)$*.*

In this chapter, we prove that given any (k,ϵ)-extractor Ext, most restricted extractors with a quasi-linear degree $\tilde{O}(\frac{n}{\epsilon^2})$ from Ext are $(k, 3\epsilon)$-extractors for a constant number of output bits, despite the degree of Ext.

Theorem 1.1.4. *There exists a universal constant* C *such that given any strong* (k,ϵ)*-extractor* $\mathsf{Ext} : \{0,1\}^n \times \{0,1\}^t \to \{0,1\}^m$*, for* $D = C \cdot \frac{n \cdot 2^m}{\epsilon^2} \cdot \log^2 \frac{n \cdot 2^m}{\epsilon}$ *random seeds* $y_1, \ldots, y_D \in \{0,1\}^t$*,* $\mathsf{Ext}_{(y_1,\ldots,y_D)}$ *is a strong* $(k, 3\epsilon)$*-extractor with probability 0.99.*

At the same time, the same statement of Theorem 1.1.4 holds for extractors. But in this case, the dependency 2^m on the degree D of restricted extractors is necessary to guarantee its error is less than $1/2$ on m output bits.

Proposition 1.1.5. *There exists a* (k,ϵ)*-extractor* $\mathsf{Ext} : \{0,1\}^n \times \{0,1\}^t \to \{0,1\}^m$ *with* $k = 1$ *and* $\epsilon = 0$ *such that any restricted extractor of* Ext *requires the degree* $D \geq 2^{m-1}$ *to guarantee its error is less than* $1/2$*.*

Though the dependency 2^m is necessary for extractors in the above proposition, it may not be necessary for *strong* extractors.

Applications of Theorem 1.1.4. In a seminal work, Impagliazzo, Levin, and Luby [ILL89] proved the Leftover Hash Lemma, i.e., all functions from an almost-universal hash family constitute a strong extractor. We first define universal hash families and almost universal hash families.

Definition 7.0.2 (Universal hash families by Carter and Wegman [CW79])**.** *Let H be a family of functions mapping $\{0,1\}^n$ to $\{0,1\}^m$. H is universal if*

$$\forall x, y \in \{0,1\}^n (x \neq y), \Pr_{h \sim H}[h(x) = h(y)] \leq 2^{-m}.$$

Moreover, we say H is almost universal if,

$$\forall x, y \in \{0,1\}^n (x \neq y), \Pr_{h \sim H}[h(x) = h(y)] \leq 2^{-m} + 2^{-n}.$$

We discuss several applications of Theorem 1.1.4 based on the Leftover Hash Lemma [ILL89] on almost universal hash families.

Lemma 7.0.3 (Leftover Hash Lemma [ILL89])**.** *For any n and m, let $\mathcal{H}$ be a family of T hash functions $\{h_1, \cdots, h_T\}$ mapping $[2^n]$ to $[2^m]$ such that for any distinct x and y, $\Pr_{h \sim H}[h(x) = h(y)] \leq 2^{-n} + 2^{-m}$. Then $\mathsf{Ext} : \{0,1\}^n \times [T] \to \{0,1\}^m$ defined as $\mathsf{Ext}(x, y) = h_y(x)$ is a strong (k, ϵ)-extractor for any k and ϵ satisfying $k \geq m + 2\log\frac{1}{\epsilon}$.*

For completeness, we provide a proof of the Leftover Hash Lemma in Appendix B.

Plugging the extractors of all linear transformations and Toeplitz matrices [ILL89] in Theorem 1.1.4, our result indicates that most extractors constituted by a quasi-linear number $\tilde{O}(\frac{n}{\epsilon^2})$ of linear transformations or Toeplitz matrices keep the nearly same parameters of the min-entropy and error, for a constant number of output bits. We treat the subset $\{0, 1, 2, \cdots, 2^n - 1\}$ the same as $\{0,1\}^n$ and $\mathbb{F}_2^n$ in this work.

Corollary 7.0.4. *There exists a universal constant C such that for any integers n, m, k, and $\epsilon > 0$ with $k \geq m + 2\log\frac{1}{\epsilon}$, $\mathsf{Ext}_{(A_1,\ldots,A_D)}$ with $D = C \cdot \frac{n \cdot 2^m}{\epsilon^2} \cdot \log^2 \frac{n \cdot 2^m}{\epsilon}$ random matrices $A_1, \ldots, A_D \in \mathbb{F}_2^{m \times n}$, mapping from $\mathbb{F}_2^n \times [D]$ to $\mathbb{F}_2^m$ as $\mathsf{Ext}_{(A_1,\ldots,A_D)}(x, i) = A_i \cdot x$, is a strong $(k, 3\epsilon)$-extractor with probability 0.99.*

Moreover, the same holds for D random Toeplitz matrices $A_1, \ldots, A_D \in \mathbb{F}_2^{n \times m}$.

Next we consider extractors from almost-universal hash families, which have efficient implementations and wide applications in practice. We describe a few examples of almost-universal hash families with efficient implementations.

1. Linear Congruential Hash by Carter and Wegman [CW79]: for any n and m, let p be a prime $> 2^n$ and $\mathcal{H}_1 = \{h_{a,b} | a, b \in \{0, 1, \cdots, p-1\}\}$ be the hash family defined as $h_{a,b}(x) = \big((ax+b) \mod p\big) \mod 2^m$ for every $x \in \{0, 1, \cdots, 2^n - 1\}$.

2. Multiplicative Universal Hash by Dietzfelbinger et al. [DHKP97] and Woelfel [Woe99]: for any n and m, let $\mathcal{H}_2 = \{h_{a,b} | a \in \{1, 3, 5, \cdots, 2^n - 1\}, b \in \{0, 1, \cdots, 2^{n-m} - 1\}$ be the hash family mapping $\{0, 1, \cdots, 2^n - 1\}$ to $\{0, 1, \cdots, 2^m - 1\}$ that first calculates $ax + b$ modulo 2^n then takes the high order m bits as the hash value, i.e., $h_{a,b}(x) = \big((ax+b) \mod 2^n\big) \operatorname{div} 2^{n-m}$. In C-code, this hash function could be implemented as $h_{a,b}(x) = (a * x + b) >> (n - m)$ when $n = 64$.

3. Shift Register Hash by Vazirani [Vaz87]: let p be a prime such that 2 is a generator modulo p and $a^{(i)}$ denote the ith shift of a string $a \in \mathbb{F}_2^n$, i.e., $a^{(i)} = a_{i+1}a_{i+2} \cdots a_n a_1 \cdots a_i$. For $n = p - 1$ and any $m \leq n$, let $\mathcal{H}_3 = \{h_a | a \in \mathbb{F}_2^p\}$ be the hash family mapping $\mathbb{F}_2^n$ to $\mathbb{F}_2^m$ as $\big(\langle a, 1 \circ x\rangle, \langle a^{(1)}, 1 \circ x\rangle, \cdots, \langle a^{m-1}, 1 \circ x\rangle\big)$, where $\langle w, z\rangle$ denotes the inner product of $w, z \in \mathbb{F}_2^p$ in $\mathbb{F}_2$.

Because all these hash families are almost-universal, by the Leftover Hash Lemma [ILL89], $\mathsf{Ext}(x, y) = h_y(x)$ is a strong extractor for all hash functions h_y in one family. Plugging these extractors in Theorem 1.1.4, we obtain extractors of almost optimal degrees with efficient implementations.

Corollary 7.0.5. *Let $\mathcal{H}$ be any almost-universal hash family mapping $\{0,1\}^n$ to $\{0,1\}^m$. There exists a universal constant C such that for any integer k and $\epsilon > 0$ with $k \geq m + 2\log\frac{1}{\epsilon}$, $\mathsf{Ext}_{(h_1,\ldots,h_D)}$ with $D = C \cdot \frac{n \cdot 2^m}{\epsilon^2} \cdot \log^2 \frac{n \cdot 2^m}{\epsilon}$ random hash functions $h_1, \cdots, h_D \sim \mathcal{H}$, defined as $\mathsf{Ext}_{(h_1,\cdots,h_D)}(x, i) = h_i(x)$, is a strong $(k, 3\epsilon)$-extractor with probability 0.99.*

Organization. In the rest of this chapter, we provide a few basic tools and Lemmas in Section 7.1. Then we prove Theorem 1.1.4 for extractors and Proposition 1.1.5 in Section 7.2. Finally, we prove Theorem 1.1.4 for strong extractors in Section 7.3.

7.1 Tools

Given a subset $\Lambda \in \{0,1\}^n$, we consider the flat random source of the uniform distribution over Λ, whose min-entropy is $-\log_2 \frac{1}{|\Lambda|} = \log_2 |\Lambda|$. Because any random source with min-entropy k is a linear combination of flat random sources of min-entropy k, we focus on flat random sources in the rest of this work.

We always use $N(0,1)$ to denote the standard Gaussian random variable and use the following concentration bound on Gaussian random variables [LT91].

Lemma 7.1.1. *Given any n Gaussian random variables $G_1, \cdots, G_n$ (not necessarily independent) where each G_i has expectation 0 and variance σ_i^2,*

$$\mathbb{E}\left[\max_{i\in[n]} |G_i|\right] \lesssim \sqrt{\log n} \cdot \max_{i\in[n]} \{\sigma_i\}.$$

Let S be a subset of events, X a random variable, and f any function from $S \times \mathrm{supp}(X)$ to $\mathbb{R}^+$. We state the standard symmetrization and Gaussianization [LT91, RW14] that transform bounding $\max_{\Lambda\in S} \sum_{j=1}^n f(\Lambda, x_i)$ of n independent random variables $x_1, \cdots, x_n \sim X$ to a Gaussian process.

Theorem 7.1.2. *For any integer n and n independent random samples $x_1, \cdots, x_n$ from X,*

$$\mathop{\mathbb{E}}_{x_1\sim X,\cdots,x_n\sim X}\left[\max_{\Lambda\in S}\sum_{j=1}^n f(\Lambda, x_i)\right] \le \max_{\Lambda\in S}\mathop{\mathbb{E}}_{x}\left[\sum_{j=1}^n f(\Lambda, x_j)\right] + \sqrt{2\pi}\cdot\mathop{\mathbb{E}}_{x}\left[\mathop{\mathbb{E}}_{g\sim N(0,1)^n}\left[\max_{\Lambda\in S}\left|\sum_{j=1}^n f(\Lambda, x_j) g_j\right|\right]\right].$$

The first term $\max_{\Lambda\in S}\mathop{\mathbb{E}}_{x}\left[\sum_{j=1}^n f(\Lambda, x_j)\right]$ is the largest expectation over all events Λ, and the second term is to bound the deviation of every event from its expectation simultaneously. For completeness, we provide a proof of Theorem 7.1.2 in Appendix B.

We state the Beck-Fiala theorem in the discrepancy theory [Cha00].

Theorem 7.1.3. *[Beck-Fiala theorem] Given a universe $[n]$ and a collection of subsets $S_1, \cdots, S_l$ such that each element $i \in [n]$ appears in at most d subsets, there exists an assignment $\chi : [n] \to \{\pm 1\}$ such that for each subset S_j, $|\sum_{i \in S_j} \chi(i)| < 2d$.*

7.2 Restricted Extractors

We study restricted extractors in this section and prove Theorem 1.1.4 for extractors. The main result in this section is that most sequences of $\tilde{O}(\frac{n \cdot 2^m}{\epsilon^2})$ seeds from any given extractor constitute a restricted extractor with nearly the same parameters of min entropy and error. On the other hand, we show that for certain extractors, the degree of its restrictions is $\Omega(2^m)$ to guarantee any constant error.

We first consider the upper bound on the degree of restricted extractors for all entropy-k flat sources fooling one fixed statistical test.

Lemma 7.2.1. *Let $\mathsf{Ext} : \{0,1\}^n \times \{0,1\}^t \to \{0,1\}^m$ be an (k, ϵ)-extractor and $D = C \cdot \frac{n \cdot \log^2 \frac{n}{\epsilon}}{\epsilon^2}$ for a universal constant C. Given any subset $T \subseteq \{0,1\}^m$, for D independently random seeds $y_1, \cdots, y_D$ in $\{0,1\}^t$,*

$$\mathop{\mathbb{E}}_{y_1, \cdots, y_D} \left[\max_{\Lambda : |\Lambda| = 2^k} \sum_{i=1}^{D} \Pr\left[\mathsf{Ext}(\Lambda, y_i) \in T \right] \right] \leq D \cdot \left(\frac{|T|}{2^m} + 2\epsilon \right). \tag{7.1}$$

Proof. We symmetrize and Gaussianize the L.H.S. of (7.1) by Theorem 7.1.2:

$$\mathop{\mathbb{E}}_{y_1, \cdots, y_D} \left[\max_{\Lambda : |\Lambda| = 2^k} \sum_{i=1}^{D} \Pr\left[\mathsf{Ext}(\Lambda, y_i) \in T \right] \right] \tag{7.2}$$

$$\leq \max_{\Lambda} \mathop{\mathbb{E}}_{y} \left[\sum_{i=1}^{D} \Pr\left[\mathsf{Ext}(\Lambda, y_i) \in T \right] \right] + \sqrt{2\pi} \mathop{\mathbb{E}}_{y} \left[\mathop{\mathbb{E}}_{g \sim N(0,1)^D} \left[\max_{\Lambda} \left| \sum_{i=1}^{D} \Pr\left[\mathsf{Ext}(\Lambda, y_i) \in T \right] g_i \right| \right] \right]. \tag{7.3}$$

Because Ext is an extractor for entropy k sources with error ϵ, the first term

$$\max_{|\Lambda|=2^k} \mathbb{E}_{y_1,\cdots,y_D}[\sum_{i=1}^{D} \Pr\left[\mathsf{Ext}(\Lambda, y_i) \in T\right]] \leq D \cdot (\frac{|T|}{2^m} + \epsilon).$$

The technical result is a bound on the Gaussian process for any $y_1, \cdots, y_D$.

Claim 7.2.2. *For any $y_1, \cdots, y_D$, $\mathbb{E}_{g\sim N(0,1)^D}\left[\max_{\Lambda:|\Lambda|=2^k}\left|\sum_{i=1}^{D} \Pr\left[\mathsf{Ext}(\Lambda, y_i) \in T\right] \cdot g_i\right|\right] \leq C_0 \cdot \sqrt{nD} \cdot \log D$ for some universal constant C_0.*

We defer the proof of Claim 7.2.2 to Section 7.2.1.

By choosing the constant C large enough, for $D = C\frac{n \log^2 \frac{n}{\epsilon}}{\epsilon^2}$, we can ensure that $C_0\sqrt{nD} \cdot \log D \leq \frac{\epsilon D}{5}$. This bounds (7.3) by $D \cdot (\frac{|T|}{2^m} + \epsilon) + \epsilon D$. □

Next, we show that a restricted extractor is good with high probability. To do this, we provide a concentration bound on $\max_{|\Lambda|=2^k}\left\{\sum_{i=1}^{D} \Pr\left[\mathsf{Ext}(\Lambda, y_i) \in T\right]\right\}$. We prove that a restricted extractor with D random seeds achieves the guarantee in Lemma 7.2.1 with probability $1 - \delta$ after enlarging D by a factor of $\tilde{O}(\log \frac{1}{\delta})$.

Lemma 7.2.3. *For any $\delta > 0$, let $D = C' \cdot \frac{n \cdot \log \frac{1}{\delta}}{\epsilon^2} \cdot \log^2 \frac{n \cdot \log \frac{1}{\delta}}{\epsilon}$ for a universal constant C'. Given any (k, ϵ)-extractor $\mathsf{Ext} : \{0,1\}^n \times \{0,1\}^t \to \{0,1\}^m$ and any subset $T \subseteq \{0,1\}^m$, for D independently random seeds $y_1, \cdots, y_D$ in $\{0,1\}^t$,*

$$\Pr_{y_1,\cdots,y_D}\left[\max_{\Lambda:|\Lambda|=2^k}\left\{\sum_{i=1}^{D} \Pr\left[\mathsf{Ext}(\Lambda, y_i) \in T\right] - \frac{D \cdot |T|}{2^m}\right\} \leq D \cdot 3\epsilon\right] \geq 1 - \delta.$$

We defer the proof of Lemma 7.2.3 to Section 7.2.2. Finally we state the result about extractors.

Theorem 7.2.4. *Let $\mathsf{Ext} : \{0,1\}^n \times \{0,1\}^t \to \{0,1\}^m$ be a (k, ϵ)-extractor and $D = C \cdot \frac{n \cdot (\log \frac{1}{\delta} + 2^m)}{\epsilon^2} \cdot \log^2 \frac{n \cdot (\log \frac{1}{\delta} + 2^m)}{\epsilon}$ for a universal constant C. For a random sequence $(y_1, \cdots, y_D)$ where each $y_i \sim \{0,1\}^t$, the restricted extractor $\mathsf{Ext}_{(y_1,\cdots,y_D)}$ is a $(k, 2\epsilon)$-extractor with probability $1 - \delta$.*

Proof. We choose the error probability to be $\frac{\delta}{2^{2^m}}$ in Lemma 7.2.3 and apply a union bound over all possible statistical tests T in $\{0,1\}^m$. □

For extractors, we show that 2^m dependence in the degree is necessary.

Claim 7.2.5. *There exists a $(k=1,\epsilon=0)$-extractor Ext such that for any constant $\epsilon' \le 1/2$ and $k'>0$, any restriction $\mathsf{Ext}_{(y_1,\cdots,y_D)}$ requires $D=\Omega(2^m)$ to be an (k',ϵ')-extractor.*

Proof. Let us consider the extractor $\mathsf{Ext}:\{0,1\}^n\times\{0,1\}^m\to\{0,1\}^m$ defined as $\mathsf{Ext}(x,y)=y$. From the definition, it is an $(k=1,\epsilon=0)$-extractor. On the other hand, $\mathsf{Ext}_(y_1,\cdots,y_D)$ is an $(k',0.5)$-extractor only if $D\ge 0.5\cdot 2^m$. □

However, this lower bound may not be necessary for *strong* extractors.

7.2.1 The Chaining Argument Fooling one test

Given $y_1,\cdots,y_D$ and T, for any subset Λ, we use $\vec{p}(\Lambda)$ to denote the vector

$$\left(\Pr\left[\mathsf{Ext}(\Lambda,y_1)\in T\right],\cdots,\Pr\left[\mathsf{Ext}(\Lambda,y_D)\in T\right]\right).$$

Let $t=10$ be a fixed parameter in this proof.

We rewrite the Gaussian process

$$\mathop{\mathbb{E}}_{g\sim N(0,1)^D}\left[\max_{\Lambda\in\binom{\{0,1\}^n}{2^k}}\left|\sum_{i=1}^{D}\Pr\left[\mathsf{Ext}(\Lambda,y_i)\in T\right]\cdot g_i\right|\right]=\mathop{\mathbb{E}}_{g\sim N(0,1)^D}\left[\max_{\Lambda\in\binom{\{0,1\}^n}{2^k}}\left|\langle\vec{p}(\Lambda),g\rangle\right|\right].$$

We construct a sequence of subsets $\mathcal{F}_{t-1},\mathcal{F}_t,\cdots,\mathcal{F}_k$ of vectors in $\mathbb{R}^D$ and a sequence of maps $\pi_j:\binom{\{0,1\}^n}{2^k}\to\mathcal{F}_j$ for each j from $t-1$ to k. We first set $\mathcal{F}_k$ to be the subset of all vectors in the Gaussian process, i.e., $\mathcal{F}_k=\left\{\vec{p}(\Lambda)\middle|\Lambda\in\binom{\{0,1\}^m}{2^k}\right\}$ and $\pi_k(\Lambda)=\vec{p}(\Lambda)$. For convenience, we set $\mathcal{F}_{t-1}=\{\vec{0}\}$ and $\pi_{t-1}(\Lambda)=\vec{0}$ for any $\Lambda\in\binom{\{0,1\}^n}{2^k}$ and specify $\mathcal{F}_j$

and π_j for $j \in [t, k-1]$ later. For any $\vec{p}(\Lambda)$ in the Gaussian process, we use the equation $\vec{p}(\Lambda) = \sum_{j=k}^{t} \pi_j(\Lambda) - \pi_{j-1}(\Lambda)$ to rewrite it:

$$\mathop{\mathbb{E}}_{g\sim N(0,1)^n}\left[\max_{\Lambda\in\binom{\{0,1\}^n}{2^k}} \left|\langle \vec{p}(\Lambda), g\rangle\right|\right] = \mathop{\mathbb{E}}_{g}\left[\max_{\Lambda\in\binom{\{0,1\}^n}{2^k}} \left|\left\langle \sum_{j=k}^{t} \pi_j(\Lambda) - \pi_{j-1}(\Lambda), g\right\rangle\right|\right] \tag{7.4}$$

$$= \mathop{\mathbb{E}}_{g}\left[\max_{\Lambda\in\binom{\{0,1\}^n}{2^k}} \sum_{j=k}^{t} \left|\left\langle \pi_j(\Lambda) - \pi_{j-1}(\Lambda), g\right\rangle\right|\right] \tag{7.5}$$

$$= \sum_{j=k}^{t} \mathop{\mathbb{E}}_{g}\left[\max_{\Lambda\in\binom{\{0,1\}^n}{2^k}} \left|\left\langle \pi_j(\Lambda) - \pi_{j-1}(\Lambda), g\right\rangle\right|\right] \tag{7.6}$$

$$\lesssim \sum_{j=k}^{t} \sqrt{\log(|\mathcal{F}_j|\cdot|\mathcal{F}_{j-1}|)} \cdot \max_{\Lambda\in\binom{\{0,1\}^n}{2^k}} \left\|\pi_j(\Lambda) - \pi_{j-1}(\Lambda)\right\|_2. \tag{7.7}$$

Here (7.7) follows from the union bound over Gaussian random variables — Lemma 7.1.1. In the rest of this proof, we provide upper bounds on $\left\|\pi_j(\Lambda) - \pi_{j-1}(\Lambda)\right\|_2$ and $|\mathcal{F}_j|$ to finish the calculation of (7.7).

Two upper bounds for $\left\|\pi_j(\Lambda) - \pi_{j-1}(\Lambda)\right\|_2$. Next, we provide two methods to bound $\left\|\pi_j(\Lambda) - \pi_{j-1}(\Lambda)\right\|_2$ in (7.7). In this proof, for any map π_j and Λ, we always choose the map $\pi_j(\Lambda) = p(\Lambda')$ for some subset Λ'.

Claim 7.2.6. *Given $|\Lambda_0| \geq D^2$, there always exists $\Lambda_1 \subseteq \Lambda_0$ with size $|\Lambda_1| \in \left[|\Lambda_0|/2 - 2D, |\Lambda_0|/2 + 2D\right]$ such that*

$$\|\vec{p}(\Lambda_0) - \vec{p}(\Lambda_1)\|_2 \leq 6D^{1.5}/|\Lambda_0|.$$

Proof. We plan to use the Beck-Fiala Theorem 7.1.3 to find Λ_1 given Λ_0. Let the ground set be Λ_0 and the collection of subsets be $S_i = \left\{\alpha \in \Lambda_0 \middle| \mathsf{Ext}(\alpha, y_i) \in T\right\}$ for each $i \in [D]$

and $S_{D+1} = \Lambda_0$. Because the degree is at most $D+1$, Theorem 7.1.3 implies an assignment $\chi : \Lambda_0 \to \{\pm 1\}$ satisfying that for each S_i, $|\sum_{\alpha \in S_i} \chi(\alpha)| < 2(D+1)$. We set $\Lambda_1 = \left\{ \alpha \in \Lambda_0 \middle| \chi(\alpha) = 1 \right\}$.

Because $|\sum_{\alpha \in \Lambda_0} \chi(\alpha)| < 2(D+1)$, $\left|\Lambda_1 - \frac{|\Lambda_0|}{2}\right| < (D+1) \leq 2D$. At the same time, for each $i \in [D]$ and S_i, $\left| \left\{ \alpha \in \Lambda_1 \middle| \mathsf{Ext}(\alpha, y_i) \in T \right\} \right| - \frac{|S_i|}{2} < (D+1)$.

To finish the proof, we prove $\left| \Pr[\mathsf{Ext}(\Lambda_0, y_i) \in T] - \Pr[\mathsf{Ext}(\Lambda_1, y_i) \in T] \right| \leq \frac{6D}{|\Lambda_0|}$.

$$
\begin{aligned}
&\left| \Pr[\mathsf{Ext}(\Lambda_0, y_i) \in T] - \Pr[\mathsf{Ext}(\Lambda_1, y_i) \in T] \right| \\
=& \left| \frac{|S_i|}{|\Lambda_0|} - \frac{\left| \left\{ \alpha \in \Lambda_1 \middle| \mathsf{Ext}(\alpha, y_i) \in T \right\} \right|}{|\Lambda_1|} \right| \\
\leq& \left| \frac{|S_i|}{|\Lambda_0|} - \frac{\left| \left\{ \alpha \in \Lambda_1 \middle| \mathsf{Ext}(\alpha, y_i) \in T \right\} \right|}{|\Lambda_0|/2} \right| + \left| \frac{\left| \left\{ \alpha \in \Lambda_1 \middle| \mathsf{Ext}(\alpha, y_i) \in T \right\} \right|}{|\Lambda_0|/2} - \frac{\left| \left\{ \alpha \in \Lambda_1 \middle| \mathsf{Ext}(\alpha, y_i) \in T \right\} \right|}{|\Lambda_1|} \right| \\
<& \frac{2(D+1)}{|\Lambda_0|} + \left| \left\{ \alpha \in \Lambda_1 \middle| \mathsf{Ext}(\alpha, y_i) \in T \right\} \right| \cdot \frac{\left| |\Lambda_0|/2 - |\Lambda_1| \right|}{|\Lambda_0|/2 \cdot |\Lambda_1|} \\
<& \frac{2(D+1)}{|\Lambda_0|} + \frac{(D+1)}{|\Lambda_0|/2} \leq \frac{6D}{|\Lambda_0|}.
\end{aligned}
$$

From the definition of $\vec{p}(\Lambda_0) = \left(\Pr\left[\mathsf{Ext}(\Lambda_0, y_1) \in T\right], \cdots, \Pr\left[\mathsf{Ext}(\Lambda_0, y_D) \in T\right] \right)$, this implies $\|\vec{p}(\Lambda_0) - \vec{p}(\Lambda_1)\|_2 \leq 6D^{1.5}/|\Lambda_0|$. □

Next we provide an alternative bound for Λ_0 with a small size using the probabilistic method.

Claim 7.2.7. *Given any Λ_0 of size at least 100, there always exists $\Lambda_1 \subseteq \Lambda_0$ with size $|\Lambda_1| \in \left[|\Lambda_0|/2 - \sqrt{|\Lambda_0|}, |\Lambda_0|/2 + \sqrt{|\Lambda_0|}\right]$ such that*

$$\|\vec{p}(\Lambda_0) - \vec{p}(\Lambda_1)\|_2 \leq 6\sqrt{D/|\Lambda_0|}.$$

Proof. We first show the existence of Λ_1 with the following two properties:

1. $|\Lambda_1| \in \left[|\Lambda_0|/2 - \sqrt{|\Lambda_0|}, |\Lambda_0|/2 + \sqrt{|\Lambda_0|}\right]$.
2. $\sum_{i\in[D]} \left(\left|\left\{\alpha \in \Lambda_1 \middle| \mathsf{Ext}(\alpha, y_i) \in T\right\}\right| - \left|\left\{\alpha \in \Lambda_0 \middle| \mathsf{Ext}(\alpha, y_i) \in T\right\}\right|/2\right)^2 \leq D \cdot |\Lambda_0|$.

We pick each element $\alpha \in \Lambda_0$ to Λ_1 randomly and independently with probability $1/2$. For the first property, $\mathop{\mathbb{E}}_{\Lambda_1}\left[\left(|\Lambda_1| - |\Lambda_0|/2\right)^2\right] = |\Lambda_0|/4$ implies it holds with probability at least $3/4$.

At the same time,

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\Lambda_1}\left[\sum_{i\in[D]} \left(\left|\left\{\alpha \in \Lambda_1 \middle| \mathsf{Ext}(\alpha, y_i) \in T\right\}\right| - \left|\left\{\alpha \in \Lambda_0 \middle| \mathsf{Ext}(\alpha, y_i) \in T\right\}\right|/2\right)^2\right] \\
&= \sum_{i\in[D]} \left|\left\{\alpha \in \Lambda_0 \middle| \mathsf{Ext}(\alpha, y_i) \in T\right\}\right|/4
\end{aligned}$$

implies the second property holds with probability at least $3/4$. Therefore there exists Λ_1 satisfying both properties.

Now let us bound $\|\vec{p}(\Lambda_0) - \vec{p}(\Lambda_1)\|_2$:

$$\begin{aligned}
&\sum_{i\in[D]} \left(\Pr[\mathsf{Ext}(\Lambda_0, y_i) \in T] - \Pr[\mathsf{Ext}(\Lambda_1, y_i) \in T]\right)^2 \\
\leq& 2\sum_{i\in[D]} \left(\frac{\left|\left\{\alpha \in \Lambda_0 \middle| \mathsf{Ext}(\alpha, y_i) \in T\right\}\right|}{|\Lambda_0|} - \frac{\left|\left\{\alpha \in \Lambda_1 \middle| \mathsf{Ext}(\alpha, y_i) \in T\right\}\right|}{|\Lambda_0|/2} \right)^2 \\
&+ 2\sum_{i\in[D]} \left(\frac{\left|\left\{\alpha \in \Lambda_1 \middle| \mathsf{Ext}(\alpha, y_i) \in T\right\}\right|}{|\Lambda_0|/2} - \frac{\left|\left\{\alpha \in \Lambda_1 \middle| \mathsf{Ext}(\alpha, y_i) \in T\right\}\right|}{|\Lambda_1|} \right)^2 \\
\leq& \frac{8}{|\Lambda_0|^2} \sum_{i\in[D]} \left(\left|\left\{\alpha \in \Lambda_0 \middle| \mathsf{Ext}(\alpha, y_i) \in T\right\}\right|/2 - \left|\left\{\alpha \in \Lambda_1 \middle| \mathsf{Ext}(\alpha, y_i) \in T\right\}\right| \right)^2 \\
&+ 2\sum_{i\in[D]} \left|\left\{\alpha \in \Lambda_1 \middle| \mathsf{Ext}(\alpha, y_i) \in T\right\}\right|^2 \cdot \left(\frac{|\Lambda_1| - |\Lambda_0|/2}{|\Lambda_1| \cdot |\Lambda_0|/2} \right)^2 \\
\leq& \frac{8D}{|\Lambda_0|} + 2\sum_{i\in[D]} \frac{|\Lambda_0|}{|\Lambda_0|^2/4} \leq 16D/|\Lambda_0|.
\end{aligned}$$

□

Constructions of $\mathcal{F}_j$. We construct $\mathcal{F}_{k-1}, \cdots, \mathcal{F}_t$ to fit in with Claim 7.2.6 and 7.2.7. We define two parameters $s(j)_l$ and $s(j)_u$ on the order of 2^j for each $\mathcal{F}_j$ such that

$$\mathcal{F}_j = \left\{ \vec{p}(\Lambda) \middle| \Lambda \in \binom{\{0,1\}^n}{s(j)_l} \cup \binom{\{0,1\}^n}{s(j)_l + 1} \cup \cdots \cup \binom{\{0,1\}^n}{s(j)_u} \right\}.$$

We start with $s(k)_l = s(k)_u = 2^k$ and define $s(j)_l$ and $s(j)_u$ from $j = k-1$ to t.

1. $j > 2\log D + 8$: we define $s(j)_l = \frac{s(j+1)_l}{2} - 2D$ and $s(j)_u = \frac{s(j+1)_u}{2} + 2D$. In this proof, we bound $2^j - 4D \leq s(j)_l \leq s(j)_u \leq 2^j + 4D$.

2. $j \le 2\log D+8$: we define $s(j)_l = \frac{s(j+1)_l}{2} - \sqrt{s(j+1)_l}$ and $s(j)_u = \frac{s(j+1)_u}{2} + \sqrt{s(j+1)_u}$. We bound $0.8{\cdot}2^j \le s(j)_l \le s(j)_u \le 1.4{\cdot}2^j$ by induction. The base case is $j > 2\log D+8$, which is proved above. Because $2D$ is always less than $\sqrt{s(j+1)_l}$ for $j > 2\log D + 8$,

$$\frac{s(j)_l}{2^j} = \prod_{i=k-1}^{j} \frac{2s(i)_l}{s(i+1)_l} \ge \prod_{i=k-1}^{j} (1 - \frac{2}{\sqrt{s(i+1)_l}}) \ge 1 - \sum_{i=k-1}^{j} \frac{2}{\sqrt{s(i+1)_l}}.$$

By induction, $\sum_{i=k-1}^{j} \frac{2}{\sqrt{s(i+1)_l}} \le \sum_{i=k-1}^{t} \frac{2}{\sqrt{0.8{\cdot}2^j}} \le 0.2$ given $t = 10$. Similarly,

$$\frac{s(j)_u}{2^j} = \prod_{i=k-1}^{j} \frac{2s(i)_u}{s(i+1)_u} \le \prod_{i=k-1}^{j} (1 + \frac{2}{\sqrt{s(i+1)_u}}).$$

By induction, $\sum_{i=k-1}^{j} \frac{2}{\sqrt{s(i+1)_u}} \le \sum_{i=k-1}^{t} \frac{2}{\sqrt{1.4{\cdot}2^j}} \le 0.2$ given $t = 10$, which implies $\frac{s(j)_u}{2^j} \le 1.4$.

Constructions of π_j. Next we define π_j from $j = k$ to $j = t$ by induction. The base case is $j = k$ such that $\pi_j(\Lambda) = \vec{p}(\Lambda)$ for any Λ of size 2^k. Given Λ and $\pi_j(\Lambda) \in \mathcal{F}_j$, we define $\pi_{j-1}(\Lambda)$ using Claim 7.2.6 or 7.2.7. From the definition of $\mathcal{F}_j$, $\pi_j(\Lambda) = \vec{p}(\Lambda_j)$ for some Λ_j with size in $[s(j)_l, s(j)_u]$.

For $j > 2\log D + 8$, we apply Claim 7.2.6 on Λ_j to find Λ_{j-1} of size $|\Lambda_{j-1}| \in [|\Lambda_j| - 2D, |\Lambda_j| + 2D]$ satisfying $\|\vec{p}(\Lambda_j) - \vec{p}(\Lambda_{j-1})\|_2 \le 6D^{1.5}/|\Lambda_j|$.

For $j \le 2\log D + 8$, we apply Claim 7.2.7 on Λ_j to find Λ_{j-1} of size $|\Lambda_{j-1}| \in [|\Lambda_j| - \sqrt{|\Lambda_j|}, |\Lambda_j| + \sqrt{|\Lambda_j|}]$ satisfying $\|\vec{p}(\Lambda_j) - \vec{p}(\Lambda_{j-1})\|_2 \le 6\sqrt{D/|\Lambda_j|}$.

Thus $|\Lambda_{j-1}|$ is always in $[s(j-1)_l, s(j-1)_u]$, which indicates $\vec{p}(\Lambda_{j-1})$ is in $\mathcal{F}_{j-1}$. We set $\pi_{j-1}(\Lambda) = \vec{p}(\Lambda_{j-1})$.

To finish this proof, we plug $0.8{\cdot}2^j \le s(j)_l \le s(j)_u \le 1.4{\cdot}2^j$ and $|\mathcal{F}_j| = \sum_{i=s(j)_l}^{s(j)_u} \binom{2^n}{i} \le$

$(s(j)_u - s(j)_l + 1) \cdot \binom{2^n}{s(j)_u} \le 2^j \cdot 2^{n \cdot 2^j} \le 2^{2n \cdot 2^j}$ into (7.7).

$$
\begin{aligned}
&\sum_{j=k}^{t} \sqrt{\log(|\mathcal{F}_j| \cdot |\mathcal{F}_{j-1}|)} \cdot \max_{\Lambda \in \binom{\{0,1\}^n}{2^k}} \left\| \pi_j(\Lambda) - \pi_{j-1}(\Lambda) \right\|_2 \\
\le &\sum_{j=k}^{2\log D+9} \sqrt{4n \cdot 2^j} \cdot 6D^{1.5}/s(j)_l + \sum_{j=2\log D+8}^{t} \sqrt{4n \cdot 2^j} \cdot 6\sqrt{D/s(j)_l} \\
\lesssim &\sum_{j=k}^{2\log D+9} \sqrt{n \cdot 2^j} \cdot D^{1.5}/2^j + \sum_{j=2\log D+8}^{t} \sqrt{n \cdot 2^j} \cdot \sqrt{D/2^j} \\
\le &\sum_{j=k}^{2\log D+9} \sqrt{nD} \cdot \frac{D}{2^{j/2}} + \sum_{j=2\log D+8}^{t} \sqrt{nD} \\
\lesssim &\log D \cdot \sqrt{nD}.
\end{aligned}
$$

7.2.2 Larger Degree with High Confidence

We finish the proof of Lemma 7.2.3 in this section. Given $(y_1, \cdots, y_D) \in \{0,1\}^{t \times D}$ and T, we consider the error vector in $\mathbb{R}^{\binom{2^n}{2^k}}$:

$$
\mathsf{Err}(y_1, \cdots, y_D) = \left(\sum_{i=1}^{D} \left(\Pr[\mathsf{Ext}(\Lambda, y_i) \in T] - \frac{T}{2^m} \right) \right)_{\Lambda \in \binom{\{0,1\}^n}{2^k}}.
$$

Because $\max_{\Lambda:|\Lambda|=2^k} \sum_{i=1}^{D} \left(\Pr\left[\mathsf{Ext}(\Lambda, y_i) \in T\right] - \frac{|T|}{2^m} \right) \le \|\mathsf{Err}(y_1, \cdots, y_D)\|_\infty$, we will prove

$$
\Pr_{y_1, \cdots, y_D}[\|\mathsf{Err}(y_1, \cdots, y_D)\|_\infty \ge 3\epsilon D] \le \delta \text{ for } D = C' \cdot \frac{n \cdot \log \frac{1}{\delta}}{\epsilon^2} \cdot \log^2 \frac{n \cdot \log \frac{1}{\delta}}{\epsilon}. \tag{7.8}
$$

Since $\mathsf{Err}(y_1, \cdots, y_D) = \mathsf{Err}(y_1, \emptyset, \cdots, \emptyset) + \mathsf{Err}(\emptyset, y_2, \emptyset, \cdots, \emptyset) + \cdots + \mathsf{Err}(\emptyset, \cdots, \emptyset, y_D)$, we plan to apply a concentration bound to prove (7.8).

Our main tool is a concentration inequality of Ledoux and Talagrand [LT91] for symmetric vectors. For convenience, we use the following version for any Banach space from Rudelson and Vershynin, which is stated as Theorem 3.8 in [RV08].

Theorem 7.2.8. *Given a Banach space with norm $\|\cdot\|$, let $Y_1, \cdots, Y_m$ be independent and symmetric random vectors taking values in it with $\|Y_j\| \le r$ for all $j \in [m]$. There exists an absolute constant C_1 such that for any integers $l \ge q$, and any $t > 0$, the random variable $\|\sum_{j=1}^m Y_j\|$ satisfies*

$$\Pr_{Y_1,\cdots,Y_m}\left[\|\sum_{j=1}^m Y_j\| \ge 8q\,\mathbb{E}\left[\|\sum_{j=1}^m Y_j\|\right] + 2r\cdot l + t\right] \le (\frac{C_1}{q})^l + 2\exp(-\frac{t^2}{256q\,\mathbb{E}[\|\sum_{j=1}^m Y_j\|]^2}).$$

To apply this theorem for *symmetric* random vectors, we symmetrize our goal $\mathsf{Err}(y_1, \cdots, y_D)$ as follows. Given a subset $T \subseteq \{0,1\}^m$ and $2D$ seeds $(y_1, \cdots, y_D)$ and $(z_1, \cdots, z_D)$, we define a vector $\Delta_{y,z}$ from $\binom{\{0,1\}^n}{2^k}$ to $\mathbb{R}$:

$$\Delta_{y,z}(\Lambda) = \sum_{i=1}^{D} (\Pr[\mathsf{Ext}(\Lambda, y_i) \in T] - \Pr[\mathsf{Ext}(\Lambda, z_i) \in T])$$

We use the ℓ_∞ norm in this section:

$$\|\Delta_{y,z}\|_\infty = \max_{\Lambda \in \binom{\{0,1\}^n}{2^k}} |\Delta_{y,z}(\Lambda)|.$$

Next we use the following Lemma to bridge Theorem 7.2.8 for symmetric random vectors and our goal (7.8).

Lemma 7.2.9. *When we generate $y = (y_1, \cdots, y_D)$ and $z = (z_1, \cdots, z_D)$ independently from the uniform distribution of $\{0,1\}^{D\times t}$, for $\mathsf{Err}(y)$ over the random choices $y = (y_1, \cdots, y_D)$,*

$$\Pr_y\left[\|\mathsf{Err}(y)\|_\infty \ge 2\mathop{\mathbb{E}}_y[\|\mathsf{Err}(y)\|_\infty] + \delta\right] \le 2\cdot \Pr_{y,z}[\|\Delta_{y,z}\|_\infty \ge \delta].$$

Proof. Let Z and Z' be independent identically distributed non-negative random variables. We use the following fact from [RV08]

$$\Pr\left[Z \ge 2\,\mathbb{E}[Z] + \delta\right] \le 2\Pr\left[Z - Z' \ge \delta\right]. \tag{7.9}$$

The reason is that

$$\begin{aligned}&\Pr_Z\left[Z \geq 2\,\mathbb{E}[Z] + \delta\right] \leq \Pr_{Z,Z'}\left[Z - Z' \geq \delta | Z' \leq 2\,\mathbb{E}[Z]\right]\\ \leq &\frac{\Pr\left[Z - Z' \geq \delta \wedge Z' \leq 2E[Z]\right]}{\Pr_{Z'}\left[Z' \leq 2\,\mathbb{E}[Z']\right]} \leq \frac{\Pr\left[Z - Z' \geq \delta\right]}{1/2}.\end{aligned}$$

By plugging $Z = \|\mathsf{Err}(y)\|_\infty$ in (7.9),

$$\Pr_y\left[\|\mathsf{Err}(y)\|_\infty \geq 2\mathop{\mathbb{E}}_y\left[\|\mathsf{Err}(y)\|_\infty\right] + \delta\right] \leq 2\Pr_{y,z}\left[\|\mathsf{Err}(y)\|_\infty - \|\mathsf{Err}(z)\|_\infty \geq \delta\right].$$

Then, for any $y = (y_1, \cdots, y_D)$ and $z = (z_1, \cdots, z_D)$, we bound $\|\mathsf{Err}(y)\|_\infty - \|\mathsf{Err}(z)\|_\infty$ by

$$\begin{aligned}&\max_{\Lambda \in \binom{\{0,1\}^n}{2^k}}\left\{\left|\sum_{i\in[D]}\Pr\left[\mathsf{Ext}(\Lambda, y_i) \in T\right] - \frac{D\cdot|T|}{2^m}\right|\right\} - \max_{\Lambda \in \binom{\{0,1\}^n}{2^k}}\left\{\left|\sum_{i\in[D]}\Pr\left[\mathsf{Ext}(\Lambda, z_i) \in T\right] - \frac{D\cdot|T|}{2^m}\right|\right\}\\ \leq &\max_{\Lambda \in \binom{\{0,1\}^n}{2^k}}\left\{\left|\sum_{i\in[D]}\Pr\left[\mathsf{Ext}(\Lambda, y_i) \in T\right] - \frac{D\cdot|T|}{2^m}\right| - \left|\sum_{i\in[D]}\Pr\left[\mathsf{Ext}(\Lambda, z_i) \in T\right] - \frac{D\cdot|T|}{2^m}\right|\right\}\\ \leq &\max_{\Lambda \in \binom{\{0,1\}^n}{2^k}}\left\{\left|\left(\sum_{i\in[D]}\Pr\left[\mathsf{Ext}(\Lambda, y_i) \in T\right] - \frac{D\cdot|T|}{2^m}\right) - \left(\sum_{i\in[D]}\Pr\left[\mathsf{Ext}(\Lambda, z_i) \in T\right] - \frac{D\cdot|T|}{2^m}\right)\right|\right\}\\ = &\max_{\Lambda \in \binom{\{0,1\}^n}{2^k}}\left\{\left|\sum_{i\in[D]}\Pr\left[\mathsf{Ext}(\Lambda, y_i) \in T\right] - \sum_{i\in[D]}\Pr\left[\mathsf{Ext}(\Lambda, z_i) \in T\right]\right|\right\} = \|\Delta_{y,z}\|_\infty\end{aligned}$$

From the discussion above, we have

$$\Pr_y\left[\|\mathsf{Err}(y)\|_\infty \geq 2\mathop{\mathbb{E}}_y\left[\|\mathsf{Err}(y)\|_\infty\right] + \delta\right] \leq 2\Pr_{y,z}\left[\|\Delta_{y,z}\|_\infty \geq \delta\right].$$

□

Proof of Lemma 7.2.3. From Lemma 7.2.9, it is enough to use Theorem 7.2.8 to show a concentration bound on $\Delta_{y,z}$. We first bound $\mathbb{E}[\|\mathsf{Err}_y\|_\infty]$ and $\mathbb{E}[\|\Delta_{y,z}\|_\infty]$. Notice that the

proofs of Theorem 7.1.2 and Lemma 7.2.1 indicate

$$\begin{aligned}
\mathbb{E}\left[\|\mathsf{Err}_y\|_\infty\right] =& \mathbb{E}\left[\max_{\Lambda:|\Lambda|=2^k}\left|\sum_{i=1}^{D}\left(\Pr\left[\mathsf{Ext}(\Lambda,y_i)\in T\right]-\frac{|T|}{2^m}\right)\right|\right]\\
\le& \mathbb{E}\left[\max_{\Lambda:|\Lambda|=2^k}\left\{\left|\sum_{i=1}^{D}\Pr\left[\mathsf{Ext}(\Lambda,y_i)\in T\right]-\mathop{\mathbb{E}}_{y'}\left[\sum_{i=1}^{D}\Pr\left[\mathsf{Ext}(\Lambda,y'_i)\in T\right]\right]\right|\right.\right.\\
&\left.\left.+\left|\mathop{\mathbb{E}}_{y'}\left[\sum_{i=1}^{D}\Pr\left[\mathsf{Ext}(\Lambda,y'_i)\in T\right]\right]-\frac{D\cdot|T|}{2^m}\right|\right\}\right]\\
\le&\sqrt{2\pi}\mathop{\mathbb{E}}_{y}\left[\mathop{\mathbb{E}}_{g\sim N(0,1)^n}\left[\max_{\Lambda\in\binom{\{0,1\}^n}{2^k}}\sum_{i=1}^{D}\Pr[\mathsf{Ext}(\Lambda,y_i)\in T]\cdot g_i\right]\right]+\epsilon\cdot D \quad \text{by Claim B.1}\\
\le&C_2\sqrt{nD}\cdot\log D+\epsilon D. \quad \text{by Claim 7.2.2}
\end{aligned}$$

Similarly,

$$\begin{aligned}
\mathbb{E}[\|\Delta_{y,z}\|_\infty] &= \mathop{\mathbb{E}}_{y,z}\left[\max_{\Lambda\in\binom{\{0,1\}^n}{2^k}}\left|\sum_{i=1}^{D}\left(\Pr[\mathsf{Ext}(\Lambda,y_i)\in T]-\Pr[\mathsf{Ext}(\Lambda,z_i)\in T]\right)\right|\right]\\
&\le \sqrt{2\pi}\mathop{\mathbb{E}}_{y}\left[\mathop{\mathbb{E}}_{g\sim N(0,1)^n}\left[\max_{\Lambda\in\binom{\{0,1\}^n}{2^k}}\sum_{i=1}^{D}\Pr[\mathsf{Ext}(\Lambda,y_i)\in T]\cdot g_i\right]\right] \quad \text{by Claim B.1}\\
&\le C_2\sqrt{nD}\cdot\log D \quad \text{by Claim 7.2.2}
\end{aligned}$$

Now we rewrite $\Delta_{y,z}(\Lambda)=\sum_{i=1}^{D}\left(\Pr[\mathsf{Ext}(\Lambda,y_i)\in T]-\Pr[\mathsf{Ext}(\Lambda,z_i)\in T]\right)$ as the summation of D *symmetric and independent* random variables

$$\Delta_{y_i,z_i}(\Lambda)=\left(\Pr[\mathsf{Ext}(\Lambda,y_i)\in T]-\Pr[\mathsf{Ext}(\Lambda,z_i)\in T]\right)$$

for $\Lambda\in\binom{\{0,1\}^n}{2^k}$. Next we bound each term

$$r=\max_{y_i,z_i}\left\{\|\Delta_{y_i,z_i}\|_\infty\right\}=\max_{y_i,z_i,\Lambda\in\binom{\{0,1\}^n}{2^k}}\left|\Pr[\mathsf{Ext}(\Lambda,y_i)\in T]-\Pr[\mathsf{Ext}(\Lambda,z_i)\in T]\right|\le 1.$$

We choose the parameters $q=2C_1=\Theta(1), l=\log\frac{1}{\delta}\le\epsilon D/10, t=\epsilon D/3$ and plug them in Theorem 7.2.8 to bound

$$\Pr\left[\|\Delta_{y,z}\|_\infty\ge 8q\cdot C_2\sqrt{nD}\cdot\log D+2r\cdot l+t\right]\le 2^{-l}+2e^{-\frac{t^2}{256q\,\mathbb{E}[\|\Delta_{y,z}\|_\infty]^2}}\le 3\delta$$

while $8q \cdot C_2\sqrt{nD} \cdot \log D + 2r \cdot l + t \leq 0.8\epsilon D$.

Since $\mathbb{E}[\|\mathsf{Err}(y)\|_\infty] \leq 1.1\epsilon D$, we have $\Pr[\|\mathsf{Err}(y)\|_\infty \geq 3 \cdot \epsilon D] \leq 3\delta$. □

7.3 Restricted Strong Extractors

We extend our techniques to strong extractors in this section.

Theorem 7.3.1. *Let* $D = C \cdot \frac{n \cdot 2^m}{\epsilon^2} \cdot (\log \frac{n}{\epsilon} + m)^2$ *for a universal constant* C *and* $\mathsf{Ext} : \{0,1\}^n \times \{0,1\}^t \to \{0,1\}^m$ *be any strong* (k, ϵ)*-extractor. For* D *independently random seeds* $y_1, \cdots, y_D$, $\mathsf{Ext}_{(y_1,\cdots,y_D)}$ *is a strong extractor for entropy* k *sources with expected error* 2ϵ.

Similar to Lemma 7.2.3, when we enlarge the degree by a factor of $\tilde{O}(\log \frac{1}{\delta})$, we improve the guarantee to a high probability $1 - \delta$ instead of an expected error.

Corollary 7.3.2. *For any* $\delta > 0$, *let* $D = C \cdot \frac{n \cdot 2^m \log \frac{1}{\delta}}{\epsilon^2} \cdot \log^2 \frac{n \cdot 2^m \log \frac{1}{\delta}}{\epsilon}$ *for a universal constant* C. *Given any strong* (k, ϵ)*-extractor* $\mathsf{Ext} : \{0,1\}^n \times \{0,1\}^t \to \{0,1\}^m$, *for* D *independently random seeds* $y_1, \cdots, y_D$, $\mathsf{Ext}_{(y_1,\cdots,y_D)}$ *is a strong* $(k, 3\epsilon)$*-extractor with probability at least* $1 - \delta$.

The proof of Theorem 7.3.1 follows the same outline of the proof of Lemma 7.2.1 with different parameters. We apply a chaining argument to bound the L^1 error of all entropy k sources Λ:

$$\max_{|\Lambda|=2^k} \left\{ \sum_{i=1}^{D} \left(\sum_{\alpha \in \{0,1\}^m} |\Pr[\mathsf{Ext}(\Lambda, y_i) = \alpha] - 2^{-m}| \right) \right\},$$

instead of bounding the error over all statistical tests in degree D strong extractors.

Proof of Theorem 7.3.1. We plan to show

$$\mathop{\mathbb{E}}_{y_1,\cdots,y_D} \left[\max_{\Lambda \in \binom{\{0,1\}^n}{2^k}} \sum_{i=1}^{D} \sum_{\alpha \in \{0,1\}^m} \left| \Pr_{x \sim \Lambda}[\mathsf{Ext}_0(x, y_i) = \alpha] - 2^{-m} \right| \right] \leq 4\epsilon D. \tag{7.10}$$

For convenience, we use $\Pr[\mathsf{Ext}_0(\Lambda, y_i) = \alpha]$ to denote $\Pr_{x\sim\Lambda}[\mathsf{Ext}_0(x, y_i) = \alpha]$ and $\mathsf{Err}_y(\Lambda)$ to denote the error of the seed y and subset Λ , i.e., $\mathsf{Err}_y(\Lambda) = \sum_{\alpha\in\{0,1\}^m} \left|\Pr[\mathsf{Ext}_0(\Lambda, y) = \alpha] - 2^{-m}\right|$. We use these notations to rewrite (7.10) as

$$\mathop{\mathbb{E}}_{y_1,\cdots,y_D}\left[\max_{\Lambda\in\binom{\{0,1\}^n}{2^k}}\sum_{i=1}^{D}\mathsf{Err}_{y_i}(\Lambda)\right].$$

Then we symmetrize and Gaussianize it by Theorem 7.1.2:

$$\mathop{\mathbb{E}}_{y}\left[\max_{\Lambda\in\binom{\{0,1\}^n}{2^k}}\sum_{i=1}^{D}\mathsf{Err}_{y_i}(\Lambda)\right] \le \max_{\Lambda\in\binom{\{0,1\}^n}{2^k}}\mathop{\mathbb{E}}_{y}\left[\sum_{i=1}^{D}\mathsf{Err}_{y_i}(\Lambda)\right] + \sqrt{2\pi}\mathop{\mathbb{E}}_{y}\left[\mathop{\mathbb{E}}_{g}\left[\max_{\Lambda\in\binom{\{0,1\}^n}{2^k}}\left|\sum_{i=1}^{D}\mathsf{Err}_{y_i}(\Lambda)\cdot g_i\right|\right]\right]. \tag{7.11}$$

Because Ext_0 is a strong extractor, the first term $\mathop{\mathbb{E}}_{y_1,\cdots,y_D}\left[\sum_{i=1}^{D}\mathsf{Err}_{y_i}(\Lambda)\right]$ is at most $2\epsilon D$ for any Λ of size 2^k.

To bound the second term in (7.11), we fix the seeds $y_1, \cdots, y_D$ and bound the Gaussian process.

Claim 7.3.3. *For any seeds $y_1, \cdots, y_D$, $\mathop{\mathbb{E}}_{g}\left[\max_{\Lambda\in\binom{\{0,1\}^n}{2^k}}\left|\sum_{i\in[D]}\mathsf{Err}_{y_i}(\Lambda)\cdot g_i\right|\right] \le C_0(\log D+m)\sqrt{nD\cdot 2^m}$ for a constant C_0.*

We defer the proof of this claim to Section 7.3.1. We finish the proof by bounding (7.10) as follows:

$$\begin{aligned}
&\mathop{\mathbb{E}}_{y_1,\cdots,y_D}\left[\max_{\Lambda\in\binom{\{0,1\}^n}{2^k}}\sum_{i=1}^{D}\mathsf{Err}_{y_i}(\Lambda)\right]\\
\le&\sqrt{2\pi}\cdot\mathop{\mathbb{E}}_{y_1,\cdots,y_D}\left\{\mathop{\mathbb{E}}_{g}\left[\max_{\Lambda\in\binom{\{0,1\}^n}{2^k}}\left|\sum_{i}\mathsf{Err}_{y_i}(\Lambda)\cdot g_i\right|\right]\right\} + \epsilon D\\
\le&\sqrt{2\pi}\cdot C_0(\log D + m)\sqrt{nD\cdot 2^m} + D\cdot\epsilon.
\end{aligned}$$

We choose $D = 10C_0^2\cdot\frac{n(\log\frac{n}{\epsilon}+m)^2\cdot 2^m}{\epsilon^2}$ such that

$$\mathop{\mathbb{E}}_{y_1,\cdots,y_D}\left[\max_{\Lambda\in\binom{\{0,1\}^n}{2^k}}\sum_{i=1}^{D}\mathsf{Err}_{y_i}(\Lambda)\right] \le 4\epsilon D.$$

This indicates the error of the strong linear extractor constituted by $A_1, \cdots, A_D$ is 2ϵ in statistical distance. □

Bottleneck of the chaining argument. In our proof of Theorem 7.3.1, we use the following relaxation to bound the distance of two vectors in the Gaussian process P corresponding to two subsets Λ and Λ':

$$\left\| \left(\|\mathsf{Ext}(\Lambda, y_i) - U_m\|_1 \big|_{i=1,\cdots,D} \right) - \left(\|\mathsf{Ext}(\Lambda', y_i) - U_m\|_1 \big|_{i=1,\cdots,D} \right) \right\|_2^2 \tag{7.12}$$

$$= \sum_{i=1}^{D} \left(\|\mathsf{Ext}(\Lambda, y_i) - U_m\|_1 - \|\mathsf{Ext}(\Lambda', y_i) - U_m\|_1 \right)^2 \tag{7.13}$$

$$\leq \sum_{i=1}^{D} (\|\mathsf{Ext}(\Lambda, y_i) - \mathsf{Ext}(\Lambda', y_i)\|_1)^2. \tag{7.14}$$

The shortcoming of our approach is that the subset chaining argument provides a tight analysis on (7.14) but not (7.12).

We show that the Gaussian process under the distance (7.14) is $\Omega(\sqrt{2^m})$ from the Sudakov minoration. For example, let us consider the distance of the first coordinate $\|\mathsf{Ext}(\Lambda, y_1) - \mathsf{Ext}(\Lambda', y_1)\|_1$. Because of the existence of random codes with constant rate and linear distance, there exists $l = \exp(2^m)$ subsets $T_1, \cdots, T_l$ in $\{0,1\}^m$ such that $|T_i \setminus T_j| = \Omega(2^m)$ for any $i \neq j$. Let $\Lambda_1, \cdots, \Lambda_l$ be the inverse images of $T_1, \cdots, T_l$ in $\mathsf{Ext}(\cdot, y_1)$. Then $\|\mathsf{Ext}(\Lambda_i, y_1) - \mathsf{Ext}(\Lambda_j, y_1)\|_1 = \Omega(1)$ for any $i \neq j$ from the distance of the code, which indicates the Gaussian process is $\Omega(2^m)$ from the Sudakov minoration for the distance (7.14).

7.3.1 The Chaining Argument of Strong Extractors

We prove Claim 7.3.3 in this section. We fix a parameter $t = 8$ in this proof.

Recall that $y_1, \cdots, y_D$ are fixed in this section, we use $\mathsf{Err}(\Lambda)$ to denote the vector $(\mathsf{Err}_{y_1}(\Lambda), \cdots, \mathsf{Err}_{y_D}(\Lambda))$. We rewrite the Gaussian process as

$$\mathop{\mathbb{E}}_{g}\left[\max_{\Lambda\in\binom{2^n}{2^k}}\left|\sum_{i\in[D]}\mathsf{Err}_{y_i}(\Lambda)\cdot g_i\right|\right]=\mathop{\mathbb{E}}_{g}\left[\max_{\Lambda\in\binom{2^n}{2^k}}\left|\left\langle\mathsf{Err}(\Lambda),g\right\rangle\right|\right].$$

We define a sequence of subsets $\mathcal{F}_{t-1},\mathcal{F}_t,\mathcal{F}_{t+1},\cdots,\mathcal{F}_k$ of vectors in $\mathbb{R}^D$ where $\mathcal{F}_{t-1}=\{\vec{0}\}$, $|\mathcal{F}_i|=\mathrm{poly}(\binom{2^n}{2^i})$, and $\mathcal{F}_k=\left\{\mathsf{Err}(\Lambda)\middle|\Lambda\in\binom{\{0,1\}^n}{2^k}\right\}$. For each i from t to k, we construct a map $\pi_i:\mathcal{F}_k\to\mathcal{F}_i$, except that π_k is the identity map and $\pi_{t-1}(v)=\vec{0}$ for any v. For any vector $v\in\mathcal{F}_k$,

$$v=\sum_{j=k}^{t}\pi_j(v)-\pi_{j-1}(v).$$

We plug these notations into the Gaussian process:

$$\begin{aligned}
\mathop{\mathbb{E}}_{g}\left[\max_{\Lambda\in\binom{2^n}{2^k}}\left|\left\langle\mathsf{Err}(\Lambda),g\right\rangle\right|\right]&=\mathop{\mathbb{E}}_{g}\left[\max_{v\in\mathcal{F}_k}\left|\left\langle v,g\right\rangle\right|\right] &(7.15)\\
&=\mathop{\mathbb{E}}_{g}\left[\max_{v\in\mathcal{F}_k}\left|\left\langle\sum_{j=k}^{t}\pi_j(v)-\pi_{j-1}(v),g\right\rangle\right|\right] &(7.16)\\
&\le\mathop{\mathbb{E}}_{g}\left[\max_{v\in\mathcal{F}_k}\sum_{j=k}^{t}\left|\left\langle\pi_j(v)-\pi_{j-1}(v),g\right\rangle\right|\right] &(7.17)\\
&\le\sum_{j=k}^{t}\mathop{\mathbb{E}}_{g}\left[\max_{v\in\mathcal{F}_k}\left|\left\langle\pi_j(v)-\pi_{j-1}(v),g\right\rangle\right|\right] &(7.18)\\
&\lesssim\sum_{j=k}^{t}\sqrt{\log|\mathcal{F}_j|\cdot|\mathcal{F}_{j-1}|}\cdot\max_{v}\|\pi_j(v)-\pi_{j-1}(v)\|_2. &(7.19)
\end{aligned}$$

We first construct $\mathcal{F}_j$ from $j=k$ to $j=t$ then define their maps $\pi_{k-1},\cdots,\pi_t$. To construct $\mathcal{F}_j$, we will specify two parameters $s(j)_l=s(j)_u=\Theta(2^j)$ for the size of Λ such that

$$\mathcal{F}_j=\left\{\mathsf{Err}(\Lambda)\middle|\Lambda\in\binom{\{0,1\}^n}{s(j)_l}\cup\binom{\{0,1\}^n}{s(j)_l+1}\cdots\cup\binom{\{0,1\}^n}{s(j)_u}\right\}.$$

Notice that the size of each subset $\mathcal{F}_j$ is bounded by

$$|\mathcal{F}_j| \le \binom{2^n}{s(j)_l} + \cdots + \binom{2^n}{s(j)_u}.$$

The base case is $s(k)_l = s(k)_u = 2^k$ and $\mathcal{F}_k = \big\{\mathsf{Err}(\Lambda)\big|\Lambda \in \binom{\{0,1\}^n}{2^k}\big\}$.

Construction of $\mathcal{F}_j$ for $j > 4\log D + m$: $s(j)_l = s(j+1)_l/2 - 2D$ and $s(j)_u = s(j+1)_u/2 + 2D$. We bound $s(j)_l \ge 2^j - 4D$ and $s(j)_u \le 2^j + 4D$ for all $j > 4\log D + m$.

Construction of $\mathcal{F}_j$ for $j \le 4\log D + m$: $s(j)_l = s(j+1)_l/2 - \sqrt{s(j+1)_l}$ and $s(j)_u = s(j+1)_u/2 + \sqrt{s(j+1)_u}$. We bound $s(j)_l \ge 0.8 \cdot 2^j$ because $s(t)_l/2^t = \prod_{j=k-1}^{t} \frac{2s(j)_l}{s(j+1)_l}$ is at least

$$(1-\frac{2}{\sqrt{s(t+1)_l}})\cdot(1-\frac{2}{\sqrt{s(t+2)_l}})\cdots(1-\frac{2}{\sqrt{s(k)_l}}) \ge 1-\sum_{j=t+1}^{k}\frac{2}{\sqrt{s(j)_l}} \ge 1-\sum_{j=t+1}^{k}\frac{2}{\sqrt{0.8\cdot 2^j}} \ge 0.8.$$

Similarly, we bound $s(j)_u \le 1.4 \cdot 2^j$ because

$$(1+\frac{2}{\sqrt{s(t+1)_u}})\cdot(1+\frac{2}{\sqrt{s(t+2)_u}})\cdots(1+\frac{2}{\sqrt{s(k)_u}}) \le 1+2\sum_{j=t+1}^{k}\frac{2}{\sqrt{s(j)_u}} \le 1+2\sum_{j=t+1}^{k}\frac{2}{\sqrt{1.4\cdot 2^j}} \le 1.4.$$

Construction of π_j: we construct the map π_j from $j = k-1$ to $j = t$ and bound $\|\pi_{j+1}(v) - \pi_j(v)\|_2$ for each $v \in \mathcal{F}_k$ in (7.19). We first use the Beck-Fiala Theorem in the discrepancy method to construct π_j with $j > 4\log D + m$ then use a randomized argument to construct π_j with $j \le 4\log D + m$.

Claim 7.3.4. *Given $\Lambda \ge D^4$ and D seeds $y_1, \cdots, y_D$, there always exists $\Lambda' \subseteq \Lambda$ with size $|\Lambda'| \in \big[|\Lambda|/2 - 2D, |\Lambda|/2 + 2D\big]$ such that*

$$\|\mathsf{Err}(\Lambda) - \mathsf{Err}(\Lambda')\|_2 \le 6D^{1.5}\cdot 2^m/|\Lambda|.$$

Proof. We plan to use the Beck-Fiala Theorem from the discrepancy method. We define the ground set $S = \Lambda$ and $m = 2^m \cdot D + 1$ subsets $T_1, \cdots, T_m$ to be

$$T_{(i-1)2^m+\alpha} = \left\{x \in \Lambda \middle| \mathsf{Ext}(x, y_i) = \alpha\right\} \text{ for each } \alpha \in [0, \cdots, 2^m - 1] \text{ and } i \in [D]$$

and the last $T_m = S = \Lambda$. Notice that the degree of every element $x \in \Lambda$ is $D + 1$.

From the Beck-Fiala Theorem, there always exists $\chi : \Lambda \to \{\pm 1\}$ such that

$$\text{for any } i \in [m], |\sum_{x \in T_i} \chi(x)| < 2D + 2.$$

We choose $\Lambda' = \{x | \chi(x) = 1\}$. From the guarantee of T_m, we know $|\Lambda'| \in \left[|\Lambda|/2 \pm (D+1)\right]$. Next we consider $\|\mathsf{Err}(\Lambda) - \mathsf{Err}(\Lambda')\|_2$.

We fix $\alpha \in \{0,1\}^m$ and $i \in [D]$ and bound $(\Pr[\mathsf{Ext}(\Lambda, y_i) = \alpha] - \Pr[\mathsf{Ext}(\Lambda', y_i) = \alpha])^2$ as follows.

$$\begin{aligned}
&(\Pr[\mathsf{Ext}(\Lambda, y_i) = \alpha] - \Pr[\mathsf{Ext}(\Lambda', y_i) = \alpha])^2 \\
\leq& 2\left(\Pr[\mathsf{Ext}(\Lambda, y_i) = \alpha] - \frac{|\{x \in \Lambda' | \mathsf{Ext}(x, y_i) = \alpha\}|}{|\Lambda|/2}\right)^2 \\
&+ 2\left(\frac{|\{x \in \Lambda' | \mathsf{Ext}(x, y_i) = \alpha\}|}{|\Lambda|/2} - \Pr[\mathsf{Ext}(\Lambda', y_i) = \alpha]\right)^2 \\
\leq& 2\left(\frac{|\{x \in \Lambda | \mathsf{Ext}(x, y_i) = \alpha\}| - 2|\{x \in \Lambda' | \mathsf{Ext}(x, y_i) = \alpha\}|}{|\Lambda|}\right)^2 \\
&+ 2\left(|\{x \in \Lambda' | \mathsf{Ext}(x, y_i) = \alpha\}| \cdot (\frac{1}{|\Lambda|/2} - \frac{1}{|\Lambda'|})\right)^2 \\
\leq& 2(\frac{3D}{|\Lambda|})^2 + 2\left(|\{x \in \Lambda' | \mathsf{Ext}(x, y_i) = \alpha\}| \cdot \frac{|\Lambda|/2 - |\Lambda|'}{|\Lambda|/2 \cdot |\Lambda'|}\right)^2 \\
\leq& \frac{18D^2}{|\Lambda|^2} + 2(\frac{2D}{|\Lambda|/2})^2 = 26D^2/|\Lambda|^2.
\end{aligned}$$

We bound $\|\mathsf{Err}(\Lambda) - \mathsf{Err}(\Lambda')\|_2^2$ using the above inequality.

$$\begin{aligned}
\|\mathsf{Err}(\Lambda) - \mathsf{Err}(\Lambda')\|_2^2 &= \sum_{i=1}^{D} \left(\sum_{\alpha \in \{0,1\}^m} |\Pr[\mathsf{Ext}(\Lambda, y_i) = \alpha] - 2^{-m}| - |\Pr[\mathsf{Ext}(\Lambda', y_i) = \alpha] - 2^{-m}| \right)^2 \\
&\leq \sum_{i=1}^{D} \left(\sum_{\alpha \in \{0,1\}^m} |\Pr[\mathsf{Ext}(\Lambda, y_i) = \alpha] - \Pr[\mathsf{Ext}(\Lambda', y_i) = \alpha]| \right)^2 \\
&\leq 2^m \sum_{i=1}^{D} \sum_{\alpha \in \{0,1\}^m} \left(\Pr[\mathsf{Ext}(\Lambda, y_i) = \alpha] - \Pr[\mathsf{Ext}(\Lambda', y_i) = \alpha]\right)^2 \\
&\leq 26 D^3 \cdot 2^{2m} / |\Lambda|^2.
\end{aligned}$$

□

Claim 7.3.5. *Given any Λ of size at least 100, there always exists $\Lambda' \subseteq \Lambda$ with size $|\Lambda'| \in \left[|\Lambda|/2 - \sqrt{|\Lambda|}, |\Lambda|/2 + \sqrt{|\Lambda|}\right]$ such that*

$$\|\mathsf{Err}(\Lambda) - \mathsf{Err}(\Lambda')\|_2 \leq 6\sqrt{D \cdot 2^m / |\Lambda|}.$$

Proof. We show the existence of Λ' by the probabilistic method of picking each element in Λ to Λ' with probability 1/2. Because $\mathbb{E}[|\Lambda'|] = \frac{|\Lambda|}{2}$ and $\mathbb{E}[(|\Lambda'| - \frac{|\Lambda|}{2})^2] = \frac{|\Lambda|}{4}$, Λ' satisfies

$$|\Lambda'| \in \left[\frac{|\Lambda|}{2} - \sqrt{|\Lambda|}, \frac{|\Lambda|}{2} + \sqrt{|\Lambda|}\right] \text{ with probability at least } 3/4 \text{ from the Chebyshev inequality.} \tag{7.20}$$

Next we consider

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\Lambda'} \left[\sum_{i \in [D]} \sum_{\alpha \in \{0,1\}^m} \left(\left|\{x \in \Lambda' | \mathsf{Ext}(x, y_i) = \alpha\}\right| - \left|\{x \in \Lambda | \mathsf{Ext}(x, y_i) = \alpha\}\right|/2 \right)^2 \right] \\
&= \sum_{i \in [D]} \sum_{\alpha \in \{0,1\}^m} \mathop{\mathbb{E}}_{\Lambda'} \left[\left(\left|\{x \in \Lambda' | \mathsf{Ext}(x, y_i) = \alpha\}\right| - \left|\{x \in \Lambda | \mathsf{Ext}(x, y_i) = \alpha\}\right|/2 \right)^2 \right] \\
&= \sum_{i \in [D]} \sum_{\alpha \in \{0,1\}^m} \left|\{x \in \Lambda | \mathsf{Ext}(x, y_i) = \alpha\}\right|/4 = D \cdot |\Lambda|/4.
\end{aligned}$$

With probability 3/4,

$$\sum_{i\in[D]}\sum_{\alpha\in\{0,1\}^m}\left(\left|\{x\in\Lambda|\mathsf{Ext}(x,y_i)=\alpha\}\right|/2-\left|\{x\in\Lambda'|\mathsf{Ext}(x,y_i)=\alpha\}\right|\right)^2\le D\cdot|\Lambda|. \quad (7.21)$$

We set Λ' to be a subset satisfying equations (7.20) and (7.21) and consider $\|\mathsf{Err}(\Lambda)-\mathsf{Err}(\Lambda')\|_2$.

We fix $\alpha\in\{0,1\}^m$ and $i\in[D]$ and bound $(\Pr[\mathsf{Ext}(\Lambda,y_i)=\alpha]-\Pr[\mathsf{Ext}(\Lambda',y_i)=\alpha])^2$ as follows.

$$\begin{aligned}
&(\Pr[\mathsf{Ext}(\Lambda,y_i)=\alpha]-\Pr[\mathsf{Ext}(\Lambda',y_i)=\alpha])^2\\
\le&2\left(\Pr[\mathsf{Ext}(\Lambda,y_i)=\alpha]-\frac{|\{x\in\Lambda'|\mathsf{Ext}(x,y_i)=\alpha\}|}{|\Lambda|/2}\right)^2\\
&+2\left(\frac{|\{x\in\Lambda'|\mathsf{Ext}(x,y_i)=\alpha\}|}{|\Lambda|/2}-\Pr[\mathsf{Ext}(\Lambda',y_i)=\alpha]\right)^2\\
\le&2\left(\frac{|\{x\in\Lambda|\mathsf{Ext}(x,y_i)=\alpha\}|-2|\{x\in\Lambda'|\mathsf{Ext}(x,y_i)=\alpha\}|}{|\Lambda|}\right)^2\\
&+2\left(|\{x\in\Lambda'|\mathsf{Ext}(x,y_i)=\alpha\}|\cdot(\frac{1}{|\Lambda|/2}-\frac{1}{|\Lambda'|})\right)^2\\
\le&8\frac{\left(\left|\{x\in\Lambda|\mathsf{Ext}(x,y_i)=\alpha\}|/2-|\{x\in\Lambda'|\mathsf{Ext}(x,y_i)=\alpha\}\right|\right)^2}{|\Lambda|^2}+20\frac{|\{x\in\Lambda'|\mathsf{Ext}(x,y_i)=\alpha\}|}{|\Lambda|^2},
\end{aligned} \quad (*)$$

where we use the property (7.20) in the last step to bound the second term. Next we

bound $\|\mathsf{Err}(\Lambda) - \mathsf{Err}(\Lambda')\|_2^2$ base on the above inquality:

$$\begin{aligned}
\|\mathsf{Err}(\Lambda) - \mathsf{Err}(\Lambda')\|_2^2 &= \sum_{i=1}^{D} \left(\sum_{\alpha \in \{0,1\}^m} |\Pr[\mathsf{Ext}(\Lambda, y_i) = \alpha] - 2^{-m}| - |\Pr[\mathsf{Ext}(\Lambda', y_i) = \alpha] - 2^{-m}| \right)^2 \\
&\leq \sum_{i=1}^{D} \left(\sum_{\alpha \in \{0,1\}^m} |\Pr[\mathsf{Ext}(\Lambda, y_i) = \alpha] - \Pr[\mathsf{Ext}(\Lambda', y_i) = \alpha]| \right)^2 \\
&\leq 2^m \sum_{i=1}^{D} \sum_{\alpha \in \{0,1\}^m} (\Pr[\mathsf{Ext}(\Lambda, y_i) = \alpha] - \Pr[\mathsf{Ext}(\Lambda', y_i) = \alpha])^2 \qquad \text{next apply (*)} \\
&\leq 8 \cdot 2^m \sum_{i,\alpha} \frac{\left(\left| \{x \in \Lambda | \mathsf{Ext}(x, y_i) = \alpha\} \right| /2 - |\{x \in \Lambda' | \mathsf{Ext}(x, y_i) = \alpha\}| \right)^2}{|\Lambda|^2} \\
&\quad + 20 \cdot 2^m \sum_{i,\alpha} \frac{|\{x \in \Lambda' | \mathsf{Ext}(x, y_i) = \alpha\}|}{|\Lambda|^2} \\
&\leq 8 \cdot 2^m \frac{2D \cdot |\Lambda|}{|\Lambda|^2} + 20 \cdot 2^m \frac{D \cdot |\Lambda|}{|\Lambda|^2} \leq \frac{36D \cdot 2^m}{|\Lambda|}.
\end{aligned}$$

□

Now we define our map $\pi_j : \mathcal{F}_k \to \mathcal{F}_j$ from $j = k$ to t by induction. The base case π_k is the identity map. Then we define π_{j-1} given π_j.

For $j > 4 \log D + m$, given $\Lambda \in \binom{\{0,1\}^n}{2^k}$, let $v = \pi_j\big(\mathsf{Err}(\Lambda)\big)$ be the vector in $\mathcal{F}_j$. From the definition of $\mathcal{F}_j$, there exists Λ_j of size between $[s(j)_l, s(j)_u]$ such that $v = \mathsf{Err}(\Lambda_j)$. Let Λ_{j-1} be the subset satisfying the guarantee in Claim 7.3.4 for Λ_j. We set $\pi_{j-1}\big(\mathsf{Err}(\Lambda)\big) = \mathsf{Err}(\Lambda_{j-1})$.

Similarly, for $j \leq 4 \log D + m$, given $u = \mathsf{Err}(\Lambda)$ and $\pi_j(u) = \mathsf{Err}(\Lambda_j)$ for Λ of size 2^k, we define $\pi_{j-1}(u) = \mathsf{Err}(\Lambda_{j-1})$ where Λ_{j-1} is the subset satisfying the guarantee in Claim 7.3.5 for Λ_j.

To finish the calculation of (7.19), we bound $|\mathcal{F}_j|$ by

$$|\mathcal{F}_j| \leq \binom{2^n}{s(j)_l} + \cdots + \binom{2^n}{s(j)_u} \leq 2 \cdot 2^j \cdot \binom{2^n}{1.8 \cdot 2^j} \leq 2^{2n2^j}.$$

From the all discussion above, we bound the Gaussian process in (7.19) as

$$\begin{aligned}
\mathop{\mathbb{E}}_{g}\left[\max_{\Lambda\in\binom{2^n}{2^k}}\left|\left\langle |Pj(\Lambda) - 2^{-m}\cdot\vec{1}|, g\right\rangle\right|\right] \lesssim & \sum_{j=k}^{t}\sqrt{\log|\mathcal{F}_j|\cdot|\mathcal{F}_{j-1}|}\cdot\max_{v}\|\pi_j(v)-\pi_{j-1}(v)\|_2 \\
\leq & \sum_{j=k}^{4\log D+m}\sqrt{2n\cdot 2^j}\cdot(10D^{1.5}\cdot 2^{m-j}) \\
& + \sum_{j=4\log D+m}^{t}\sqrt{2n\cdot 2^j}\cdot 10\sqrt{D\cdot 2^{m-j}} \\
\lesssim & \sum_{j=k}^{4\log D+m}\sqrt{n}\cdot\frac{(D^{1.5}\cdot 2^m)}{2^{j/2}} + \sum_{j=4\log D+m}^{t}\sqrt{2nD\cdot 2^m} \\
\lesssim & \sqrt{n}\cdot\frac{(D^{1.5}\cdot 2^m)}{D^2\cdot 2^{m/2}} + (4\log D+m)\cdot\sqrt{2nD\cdot 2^m} \\
\lesssim & (4\log D+m)\cdot\sqrt{2nD\cdot 2^m}.
\end{aligned}$$

Chapter 8

Hash Functions for Multiple-Choice Schemes

We present explicit constructions of hash families that guarantee the same maximum loads as a perfectly random hash function in the multiple-choice schemes. We construct our hash families based on the hash family of Celis et al. [CRSW13] for the 1-choice scheme, which is $O(\log \log n)$-wise independent over n bins and "almost" $O(\log n)$-wise independent over a fraction of $\text{poly}(\log n)$ bins.

We first show our hash family guarantees a maximum load of $\frac{\log \log n}{\log d} + O(1)$ in the *Uniform-Greedy* scheme [ABKU99, Voc03] with d choices. We use U to denote the pool of balls and consider placing $m = O(n)$ balls into n bins here. Without loss of generality, we always assume $|U| = \text{poly}(n)$ and d is a constant at least 2 in this work.

Theorem 8.0.1 (Informal version of Theorem 8.4.1)**.** *For any $m = O(n)$, any constants c and d, there exists a hash family with $O(\log n \log \log n)$ random bits such that given any m balls in U, with probability at least $1 - n^{-c}$, the max-load of the* Uniform-Greedy *scheme with d independent choices of h is $\frac{\log \log n}{\log d} + O(1)$.*

Our hash family has an evaluation time $O\big((\log \log n)^4\big)$ in the RAM model based on the algorithm designed by Meka et al. [MRRR14] for the hash family of Celis et al. [CRSW13].

Then we show this hash family guarantees a load balancing of $\frac{\log \log n}{d \log \phi_d} + O(1)$ in the *Always-Go-Left* scheme [Voc03] with d choices. The *Always-Go-Left* scheme [Voc03] is an asymmetric allocation scheme that partitions the n bins into d groups with equal size and

uses an unfair tie-breaking mechanism. Its allocation process provides d independent choices for each ball from the d groups separately and always chooses the left-most bin with the least load for each ball. We defer the formal description of the *Always-Go-Left* scheme to Section 8.5. Notice that the constant ϕ_d in equation $\phi_d^d = 1 + \phi_d + \cdots + \phi_d^{d-1}$ satisfies $1.61 < \phi_2 < \phi_3 < \phi_4 < \cdots < \phi_d < 2$. Compared to the *Uniform-Greedy* scheme, the *Always-Go-Left* scheme [Voc03] improves the maximum load exponentially with regard to d. Even for $d = 2$, the *Always-Go-Left* scheme improves the maximum load from $\log\log n + O(1)$ to $0.7 \log\log n + O(1)$.

Theorem 8.0.2 (Informal version of Theorem 8.5.3)**.** *For any $m = O(n)$, any constants c and d, there exists a hash family with $O(\log n \log\log n)$ random bits such that given any m balls in U, with probability at least $1 - n^{-c}$, the max-load of the* Always-Go-Left *scheme with d independent choices of h is $\frac{\log\log n}{d \log \phi_d} + O(1)$.*

At the same time, from the lower bound $\frac{\log\log n}{d \log \phi_d} - O(1)$ on the maximum load of any random d-choice scheme shown by Vöcking [Voc03], the maximum load of our hash family is optimal for d-choice schemes up to the low order term of constants.

Finally, we show our hash family guarantees the same maximum load as a perfectly random hash function in the 1-choice scheme for $m = n \cdot \text{poly}(\log n)$ balls. Given $m > n \log n$ balls in U, the maximum load of the 1-choice scheme becomes $\frac{m}{n} + O(\sqrt{\log n \cdot \frac{m}{n}})$ from the Chernoff bound. For convenience, we refer to this case of $m \geq n \log n$ balls as a *heavy load*. In a recent breakthrough, Gopalan, Kane, and Meka [GKM15] designed a pseudorandom generator of seed length $O(\log n (\log\log n)^2)$ that fools the Chernoff bound within polynomial error. Hence the pseudorandom generator [GKM15] provides a hash function with $O(\log n (\log\log n)^2)$ random bits for the heavy load case. Compared to the hash function of [GKM15], we provide a simplified construction that achieves the same maximum load but only works for $m = n \cdot \text{poly}(\log n)$ balls.

Theorem 8.0.3 (Informal version of Theorem 8.6.1)**.** *For any constants c and $a \geq 1$, there exist a hash function generated by $O(\log n \log\log n)$ random bits such that for any $m = \log^a n \cdot n$ balls, with probability at least $1 - n^{-c}$, the max-load of the n bins in the 1-choice scheme with h is $\frac{m}{n} + O\left(\sqrt{\log n} \cdot \sqrt{\frac{m}{n}}\right)$.*

8.1 Preliminaries

We use U to denote the pool of balls, m to denote the numbers of balls in U, and n to denote the number of bins. We assume $m \geq n$ and n is a power of 2 in this work. We use $\mathbb{F}_p$ to denote the Galois field of size p for a prime power p.

Definition 8.1.1. *Given a prime power p, a distribution D on $\mathbb{F}_p^n$ is a δ-biased space if for any non-trivial character function χ_α in $\mathbb{F}_p^n$, $\mathop{\mathbb{E}}_{x\sim D}[\chi_\alpha(x)] \leq \delta$.*

A distribution D on $\mathbb{F}_p^n$ is a k-wise δ-biased space if for any non-trivial character function χ_α in $\mathbb{F}_p^n$ of support size at most k, $\mathop{\mathbb{E}}_{x\sim D}[\chi_\alpha(x)] \leq \delta$.

The seminal works [NN90, AGHP90] provide small-biased spaces with optimal seed length.

Lemma 8.1.2 ([NN90, AGHP90])**.** *For any prime power p and integer n, there exist explicit constructions of δ-biased spaces on $\mathbb{F}_p^n$ with seed length $O(\log \frac{pn}{\delta})$ and explicit constructions of k-wise δ-biased spaces with seed length $O(\log \frac{kp\log n}{\delta})$.*

Given two distributions D_1 and D_2 with the same support $\mathbb{F}_p^n$, we define the statistical distance to be $\|D_1 - D_2\|_1 = \sum_{x\in\mathbb{F}_p^n} |D_1(x) - D_2(x)|$. Vazirani [Vaz86] proved that small-biased spaces are close to the uniform distribution.

Lemma 8.1.3 ([Vaz86])**.** *A δ-biased space on $\mathbb{F}_p^n$ is $\delta \cdot p^{n/2}$ close to the uniform distribution in statistical distance.*

Given a subset S of size k in $[n]$, a k-wise δ-biased space on $\mathbb{F}_p^n$ is $\delta \cdot p^{k/2}$ close to the uniform distribution on S in statistical distance.

Given a distribution D on functions from U to $[n]$, D is k-wise independent if for any k elements $x_1, \ldots, x_k$ in U, $D(x_1), \ldots, D(x_k)$ is a uniform distribution on $[n]^k$. For small-biased spaces, we choose $p = n$ and the space to be $\mathbb{F}_n^{|U|}$ in Lemma 8.1.2 and summarize the discussion above.

Lemma 8.1.4. *Given k and n, a k-wise δ-biased space from U to $[n]$ is $\delta \cdot n^{k/2}$ close to the uniform distribution from U to $[n]$ on any k balls, which needs $O(\log \frac{kn \log n}{\delta})$ random bits.*

Remark 8.1.5. *In this work, we always choose $\delta \leq 1/n$ and $k = \text{poly}(\log n)$ in the small biased spaces such that the seed length is $O(\log \frac{1}{\delta})$. At the same time, we only use k-wise small-biased spaces rather than small biased spaces to improve the evaluation time from $O(\log n)$ to $O(\log \log n)^4$.*

We state the Chernoff bound in k-wise independence by Schmidt et al. in [SSS95].

Lemma 8.1.6 (Theorem 5 (I) (b) in [SSS95])**.** *If X is the sum of k-wise independent random variables, each of which is confined to the interval $[0, 1]$ with $\mu = \mathbb{E}[X]$, then for $\delta \leq 1$ and $k \geq \delta^2 \mu \cdot e^{-1/3}$,*

$$\Pr[|X - \mu| \geq \delta\mu] \leq e^{-\delta^2 \mu/3}.$$

8.2 Witness Trees

We first provide several notation and definitions in this work. Then we review the witness tree argument of Vöcking [Voc03] for the *Uniform-Greedy* scheme.

Definition 8.2.1 (*Uniform-Greedy* with d choices)**.** *The process inserts balls in any fixed order. Let $h^{(1)}, \ldots, h^{(d)}$ be d hash functions from U to $[n]$. The allocation process works as*

follows: for each ball i, the algorithm considers d bins $\{h^{(1)}(i), \ldots, h^{(d)}(i)\}$ and puts the ball i into the bin with the least load among $\{h^{(1)}(i), \ldots, h^{(d)}(i)\}$. When there are several bins with the least load, it picks an arbitrary one.

We define the height of a ball to be the height of it on the bin allocated in the above process.

Next we follow the notation of Vöcking [Voc03] to define witness trees and pruned witness trees. Given the balls and d hash functions $h^{(1)}, \ldots, h^{(d)}$ in the allocation process, we construct a symmetric witness tree for each ball in this process.

Definition 8.2.2 (Symmetric witness trees)**.** *A symmetric witness tree T with height l for a ball b is a complete d-ary tree of height l. Every node w in this tree corresponds to a ball $T(w) \in [n]$; and the root corresponds to the ball b. A ball u in T has a ball v as its ith child iff when we allocate u in the process, ball v is the top ball in the bin $h^{(i)}\big(u\big)$. Hence $v < u$ and the bin $h^{(i)}(u)$ is in the subset $\big\{h^{(1)}(v), \ldots, h^{(d)}(v)\big\}$ of $[n]$ when v is the ith child of u.*

Next we trim the repeated nodes in a witness trees such that there is no duplicate edge after the trimming.

Definition 8.2.3 (Pruned witness trees and collisions)**.** *Given a witness tree T where nodes $v_1, \ldots, v_j$ in T correspond to the same ball, let v_1 be the node among them in the most bottom level of T. Consider the following process: first remove $v_2, \ldots, v_j$ and their subtrees; then, redirect the edges of $v_2, \ldots, v_j$ from their parents to v_1 and call these edges* collisions*. Given a symmetric witness tree T, we call the new tree without repeated nodes after the above process as the pruned witness tree of T.*

We call different witness trees with the same structure but different balls a *configuration*. For example, the configuration of symmetric witness trees with *distinct* nodes is a full d-ary tree without any collision.

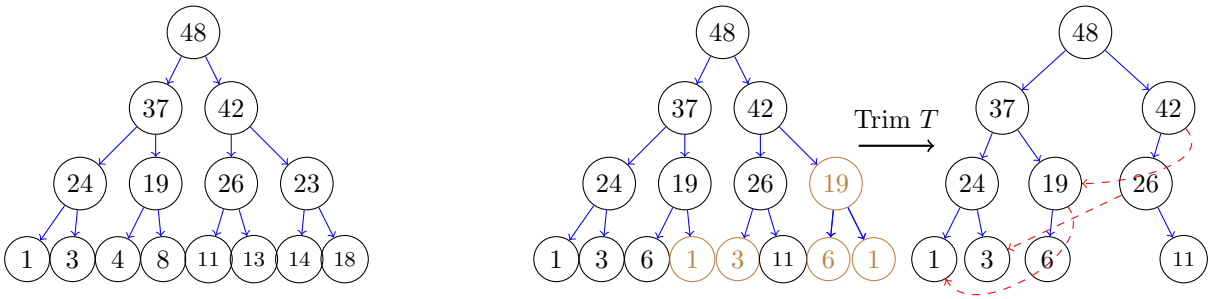

Figure 8.1: A witness tree with distinct balls and a pruned witness tree with 3 collisions

Next we define the height and size of pruned witness trees.

Definition 8.2.4 (Height of witness trees)**.** *Given any witness tree T, let the height of T be the length of the* shortest *path from the root of T to its leaves. Because $\mathit{height}(u) = \min_{v \in \mathit{children}(u)} \{\mathit{height}(v)\} + 1$, the height of the pruned witness tree equals the height of the original witness tree. Given a ball b of height h and any $h' < h$, we always consider the pruned witness tree of b with height h' whose leaves have height $h - h'$.*

At the same time, let $|T|$ denote the number of vertices in T for any *witness tree T and $|C|$ denote the number of nodes in a configuration C.*

Finally we review the argument of Vöcking [Voc03] for $m = n$ balls. One difference between this proof and Vöcking's [Voc03] proof is an alternate argument for the case of witness trees with many collisions.

Lemma 8.2.5 ([Voc03])**.** *For any constants $c \geq 2$ and d, with probability at least $1 - n^{-c}$, the max-load of the* Always-Go-Left *scheme with d independent choices from perfectly random hash functions is $\frac{\log \log n}{\log d} + O(1)$.*

Proof. We fix a parameter $l = \lceil \log_d \left((2+2c)\log n\right) + 3c + 5 \rceil$ for the height of witness trees. In this proof, we bound the probability that any symmetric witness tree of height l with

leaves of height at least 4 exists in perfectly random hash functions. From the definition of witness trees, this also bounds the probability of a ball with height $l + 4$ in the d-choice *Uniform-Greedy* scheme.

For symmetric witness trees of height l, it is sufficient to bound the probability that their pruned counterparts appear in perfectly random hash functions. We separate all pruned witness trees into two cases according to the number of edge collisions: pruned witness trees with at most $3c$ collisions and pruned witness trees with at least $3c$ collisions.

Pruned witness trees with at most $3c$ collisions. Let us fix a configuration C with at most $3c$ collisions and consider the probability any pruned witness trees with configuration C appears in perfectly random hash functions. Because each node of this configuration C corresponds a distinct ball, there are at most $n^{|C|}$ possible ways to instantiate balls into C.

Next, we fix one possible pruned witness tree T and bound the probability of the appearance of T in $h^{(1)}, \ldots, h^{(d)}$. We consider the probability of two events: every edge (u, v) in the tree T appears during the allocation process; and every leaf of T has height at least 4. For the first event, an edge (u, v) holds during the process only if the hash functions satisfy

$$h^{(i)}(u) \in \left\{h^{(1)}(v), \ldots, h^{(d)}(v)\right\}, \text{ which happens with probability at most } \frac{d}{n}. \tag{8.1}$$

Secondly, the probability that a fixed leaf ball has height at least 4 is at most 3^{-d}. A leaf ball of height 4 indicates that each bin in his choices has height at least 3. Because at most $n/3$ bins contain at least 3 balls at any moment, the probability that a random bin has height at least 3 is $\leq 1/3$. Thus the probability that d random bins have height 3 is at most 3^{-d}.

We apply a union bound on the probability that any witness tree with the configuration C appears in perfectly random hash functions:

$$n^{|C|} \cdot \prod_{(u,v)\in C} \frac{d}{n} \cdot (3^{-d})^{\text{number of leaves}} \tag{8.2}$$

We lower bound the number of edges in C by $|C|-1$ because C is connected. Next we lower bound the number of leaves. Because C is a d-ary tree with at most $3c$ collisions, the number of leaves is at least $\frac{|C|-3c}{2}$. At the same time, C is trimmed from the d-ary symmetric witness tree of height l. Thus $|C| \geq (1+d+\cdots+d^{l-3c})$. From all discussion above, we bound (8.2) by

$$n^{|C|}\cdot(\frac{d}{n})^{|C|-1}\cdot(3^{-d})^{\frac{|C|-3c}{2}} \leq n\cdot(d^{2.5}\cdot 3^{-d})^{|C|/2.5} \leq n\cdot(d^{2.5}\cdot 3^{-d})^{d^{l-3c}/2.5} \leq n\cdot(0.8)^{10(2+2c)\log n} \leq n^{-2c-1}.$$

Finally, we apply a union bound on all possible configurations with at most $3c$ collisions: the number of configurations is at most $\sum_{i=0}^{3c}(d^{l+1})^{2\cdot i} \leq n$ such that the probability of any witness tree with height l and at most $3c$ collisions existing is at most n^{-c}.

Pruned witness trees with at least $3c$ collisions. We use the extra $3c$ collisions with equation (8.1) instead of the number of leaves in this case.

Given any configuration C with at least $3c$ collisions, we consider the first $3c$ collisions $e_1,\ldots,e_{3c}$ in the BFS of C. Let C' be the induced subgraph of C that only contains nodes in $e_1,\ldots,e_{3c}$ and their ancestors in C. At the same time, the size $|C'| \leq 3c(2l+1)$ and the number of edges in C' is $|C'|+3c-1$.

Because any pruned witness tree of C exists only if its corresponding counterpart of C' exists in perfectly random hash functions, it is suffice to bound the probability of the latter event. There are at most $n^{|C'|}$ instantiations of balls in C'. For each instantiation, we bound the probability that all edges survive by (8.1):

$$(\frac{d}{n})^{\text{number of edges}} = (\frac{d}{n})^{|C'|+3c-1}.$$

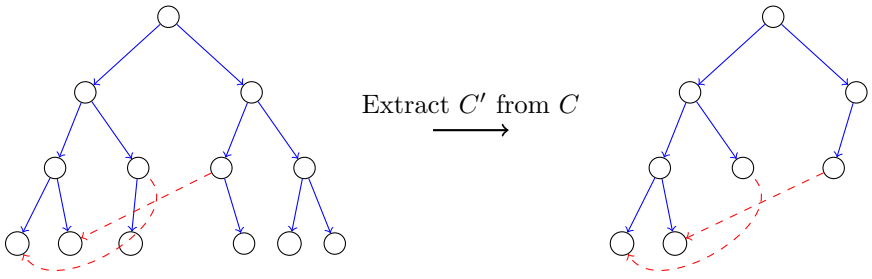

Figure 8.2: An example of extracting C' from C given two collisions.

We bound the probability that any pruned witness tree of configuration C' survives in the perfectly random hash function by

$$(\frac{d}{n})^{|C'|+3c-1} \cdot n^{|C'|} \leq (\frac{1}{n})^{3c-1} \cdot d^{(2l+2)\cdot 3c} \leq n^{-2c}.$$

Finally, we apply a union bound over all possible configurations C': there are at most $(1 + d + \cdots + d^l)^{2\cdot 3c} \leq n$ configurations of $3c$ collisions. □

Remark 8.2.6. *Because the sizes of all witness trees are bounded by* $d^{l+1} = O(\log n)$, $O_{c,d}(\log n)$*-wise independent hash functions could adopt the above argument to prove a max-load of* $\log_d \log n + O(d + c)$.

8.3 Hash Functions

We construct our hash family and show its properties for the derandomization of d-choice schemes in this section. We sketch the derandomization of Lemma 8.2.5 of Vocking's argument in Section 8.3.1.

Let $\circ$ denote the concatenation operation and $\oplus$ denote the bit-wise XOR operation.

Construction 8.3.1. *Given* $\delta_1 > 0, \delta_2 > 0$, *and two integers* k, k_g, *let*

1. $h_i : U \to [n^{2^{-i}}]$ *denote a function generated by an* $O(\log^2 n)$*-wise* δ_1*-biased space for each* $i \in [k]$*,*

2. $h_{k+1} : U \to [n^{2^{-k}}]$ *denote a function generated by an* $O(\log^2 n)$*-wise* δ_2*-biased space such that* $(h_1(x) \circ h_2(x) \circ \cdots \circ h_k(x) \circ h_{k+1}(x))$ *is a function by* U *to* $[n]$*,*

3. $g : U \to [n]$ *denote a function from a* k_g*-wise independent family from* U *to* $[n]$*.*

We define a random function $h : U \to [n]$ *in our hash family* $\mathcal{H}$ *with parameters* δ_1, δ_2, k *and* k_g *to be:*

$$h(x) = \big(h_1(x) \circ h_2(x) \circ \cdots \circ h_k(x) \circ h_{k+1}(x)\big) \oplus g(x).$$

Hence the seed length of our hash family is $O(k \log \frac{n \cdot \log^2 n \cdot \log |U|}{\delta_1} + \log \frac{n \cdot \log^2 n \cdot \log |U|}{\delta_2} + k_g \log n)$. We always choose $k \leq \log \log n, k_g = O(\log \log n), \delta_1 = 1/\text{poly}(n)$, and $\delta_2 = (\log n)^{-O(\log n)}$ such that the seed length is $O(\log n \log \log n)$.

Remark 8.3.2. *Our parameters of* $h_1 \circ \cdots \circ h_{k+1}$ *are stronger than the parameters in [CRSW13]. While the last function* h_{k+1} *of [CRSW13] is still a* δ_1*-biased space, we use* $\delta_2 = (\delta_1)^{O(k)}$ *in* h_{k+1} *to provide almost* $O(\log n)$*-wise independence on* $(\log n)^{O(\log n)}$ *subsets of size* $O(\log n)$ *for our calculations.*

Properties of h. We state the properties of h that will be used in the derandomization. Because of the k_g-wise independence in g and the $\oplus$ operation, we have the same property for h.

Property 8.3.3. *h is k_g-wise independent.*

Then we fix g and discuss $h_1 \circ \cdots \circ h_k \circ h_{k+1}$. For each $i \in [k]$, it is natural to think $h_1 \circ \cdots \circ h_i$ as a function from U to $[n^{1-\frac{1}{2^i}}]$, i.e., a hash function maps all balls into $n^{1-\frac{1}{2^i}}$ bins. Celis et al. [CRSW13] showed that for every $i \in [k]$, the number of balls in every bin of $h_1 \circ \cdots \circ h_i$ is close to its expectation $n^{\frac{1}{2^i}} \cdot \frac{m}{n}$ in $\frac{1}{\text{poly}(n)}$-biased spaces.

Lemma 8.3.4 ([CRSW13]). *Given $k = \log_2(\log n/3\log\log n)$ and $\beta = (\log n)^{-0.2}$, for any constant $c > 0$, there exists $\delta_1 = 1/\mathrm{poly}(n)$ such that given $m = O(n)$ balls, with probability at least $1 - n^{-c}$, for all $i \in [k]$, every bin in $[n^{1-\frac{1}{2^i}}]$ contains at most $(1+\beta)^i n^{\frac{1}{2^i}} \cdot \frac{m}{n}$ balls under $h_1 \circ \cdots \circ h_i$.*

For completeness, we provide a proof of Lemma 8.3.4 in Appendix C. In this work, we use the following version that after fixing g in the Construction 8.3.1, $h_1 \circ h_2 \circ \cdots \circ h_k$ still allocates the balls evenly.

Corollary 8.3.5. *For any constant $c > 0$, there exists $\delta_1 = 1/\mathrm{poly}(n)$ such that given $m = O(n)$ balls and any function $g_0 : U \to [n/\log^3 n]$, with probability at least $1 - n^{-c}$ over $h_1, \ldots, h_k$, for any bin $j \in [n^{1-\frac{1}{2^k}}] = [n/\log^3 n]$, it contains at most $1.01 \cdot \log^3 n \cdot \frac{m}{n}$ balls in the hash function $\big(h_1(x) \circ \cdots \circ h_k(x)\big) \oplus g_0(x)$.*

Next we discuss the last function h_{k+1} generated from a δ_2-biased space on $[\log^3 n]^U$. For a subset $S \subseteq U$, let $h(S)$ denote the distribution of a random function h on S and $U_{[\log^3 n]}(S)$ denote the uniform distribution over all maps from $S \to [\log^3 n]$. From Lemma 8.1.3 and the union bound, we have the following claim.

Claim 8.3.6. *Given $\delta_2 = (\log n)^{-C\log n}$, for a fixed subset S of size $\frac{C}{3} \cdot \log n$, $h_{k+1}(S)$ is $(\log n)^{-\frac{C}{2}\cdot\log n}$-close to the uniform distribution on S, i.e., $\|h_{k+1}(S) - U_{[\log^3 n]}(S)\|_1 \le (\log n)^{-\frac{C}{2}\cdot\log n}$.*

Then for $m = (\log n)^{\frac{C}{3}\cdot\log n}$ subsets $S_1, \ldots, S_m$ of size $\frac{C}{3} \cdot \log n$, we have

$$\sum_{i\in[m]} \|h_{k+1}(S_i) - U_{[\log^3 n]}(S_i)\|_1 \le m \cdot (\log n)^{-\frac{C}{2}\cdot\log n} \le (\log n)^{-\frac{C}{6}\cdot\log n}.$$

In another word, h_{k+1} is close to the uniform distribution on $(\log n)^{\frac{C}{3}\log n}$ *subsets of size* $\frac{C}{3}\log n$. However, h_{k+1} (or h) is not close to $\log n$-wise independence on n balls.

Remark 8.3.7 (Evaluation time)**.** *Our hash function has an evaluation time* $O((\log\log n)^4)$ *in the RAM model. Because we use* $(\log n)^{-O(\log\log n)}$*-biased spaces in* h_{k+1}*, we lose a factor of* $O(\log\log n)^2$ *compared to the hash family of [CRSW13]. The reason is as follows.*

g *can be evaluated by a degree* $O(\log\log n)$ *polynomial in the Galois field of size* $\mathrm{poly}(n)$*, which takes* $O(\log\log n)$ *time. The first* k *hash functions* $h_1, \ldots, h_k$ *use* $1/\mathrm{poly}(n)$*-biased spaces, which have total evaluation time* $O(k \cdot \log\log n) = O(\log\log n)^2$ *in the RAM model from [MRRR14].*

The last function h_{k+1} *in the RAM model is a* $O(\log n)$*-wise* $n^{-O(\log\log n)}$*-biased space from* U *to* $[\log^3 n]$*, which needs* $O(\log\log n)$ *words in the RAM model. Thus the evaluation time becomes* $O(\log\log n)$ *times the cost of a quadratic operation in the Galois field of size* $n^{O(\log\log n)}$*, which is* $O((\log\log n)^4)$*.*

8.3.1 Proof Overview

We sketch the derandomization of Lemma 8.2.5 in this section. Similar to the proof of Lemma 8.2.5, we bound the probability that any pruned witness tree of height $l = \log_d \log n + O(1)$ exists in $h^{(1)}, \ldots, h^{(d)}$, where each $h^{(i)} = \left(h_1^{(i)}(x) \circ \cdots \circ h_{k+1}^{(i)}(x)\right) \oplus g^{(i)}(x)$. We use the property of $h_1 \circ \cdots \circ h_{k+1}$ to derandomize the case of pruned witness trees with at most $3c$ collisions and the property of g to derandomize the other case.

Pruned witness trees with at most $3c$ **collisions.** We show how to derandomize the union bound (8.2) for a fixed configuration C with at most $3c$ collisions. There are two probabilities in (8.2): the second term $\prod_{(u,v)\in C} \frac{d}{n}$ over all edges in C and the last term $3^{-d\cdot\text{number of leaves}}$ over all leaves. We focus on the second term $\prod_{(u,v)\in C} \frac{d}{n}$ in this discussion, because it contributes a smaller probability. Since $|C| \in [d^{l-3c}, d^{l+1}] = \Theta(\log n)$, it needs $O(\log n)$-wise independence over $[n]$ bins for every possible witness trees in (8.2), which is impossible to support with $o(\log^2 n)$ bits [Sti94].

We omit $\{g^{(1)}, \ldots, g^{(d)}\}$ and focus on the other part $\left\{h_1^{(i)} \circ \cdots \circ h_k^{(i)} \circ h_{k+1}^{(i)} \middle| i \in [d]\right\}$ in this case. Our strategy is to first fix the prefixes in the d hash functions, $\left\{h_1^{(i)} \circ \cdots \circ h_k^{(i)} \middle| i \in [d]\right\}$, then recalculate (8.2) using the suffixes $h_{k+1}^{(1)}, \ldots, h_{k+1}^{(d)}$. Let T be a possible witness tree in the configuration C. For an edge (u, v) in T to satisfy (8.1), the prefixes of $h^{(1)}(v), \ldots, h^{(d)}(v)$ and $h^{(i)}(u)$ must satisfy

$$h_1^{(i)}(u) \circ \cdots \circ h_k^{(i)}(u) \in \left\{h_1^{(1)}(v) \circ \cdots \circ h_k^{(1)}(v), \ldots, h_1^{(d)}(v) \circ \cdots \circ h_k^{(d)}(v)\right\}. \tag{8.3}$$

After fixing the prefixes, let $\mathcal{F}_T$ denote the subset of possible witness trees in the configuration C that satisfy the prefix condition (8.3) for every edge. Because each bin of $[n/\log^3 n]$ receives at most $1.01 \log^3 n$ balls from every prefix function $h_1^{(j)} \circ h_2^{(j)} \circ \cdots \circ h_k^{(j)}$ by Corollary 8.3.5, we could bound

$$|\mathcal{F}_T| \leq n(d \cdot 1.01 \log^3 n)^{|C|-1} = n \cdot (1.01d)^{|C|} \cdot (\log^3 n)^{|C|-1} = (\log n)^{O(\log n)}$$

instead of $n^{|C|}$ in the original argument.

Now we consider all possible witness trees in $\mathcal{F}_T$ under the suffixes $h_{k+1}^{(1)}, \ldots, h_{k+1}^{(d)}$. We could treat $h_{k+1}^{(1)}, \ldots, h_{k+1}^{(d)}$ as $O(\log n)$-wise independent functions for all possible witness trees in $\mathcal{F}_T$ from Claim 8.3.6, because $|C| = O(\log n)$ and $|\mathcal{F}_T| = (\log n)^{O(\log n)}$. In the next step, we use $O(\log n)$-wise independence to rewrite (8.2) and finish the proof of this case.

Pruned witness trees with at least $3c$ collisions. In our alternate argument of this case in Lemma 8.2.5, the subconfiguration C' of C has at most $3c \cdot (2l + 1)$ nodes and $3c \cdot (2l + 1) + 3c$ edges. Since $l = \log_d \log n + O(1)$, the number of edges in C' is $O(\log \log n)$. By choosing $k_g = \Theta(\log \log n)$ with a sufficiently large constant, h with k_g-wise independence supports the argument in Lemma 8.2.5.

8.4 The *Uniform Greedy* Scheme

We prove our main result for the *Uniform-Greedy* scheme — Theorem 8.0.1 in this section.

Theorem 8.4.1. *For any $m = O(n)$, any constant $c \geq 2$, and integer d, there exists a hash family $\mathcal{H}$ from Construction 8.3.1 with $O(\log n \log\log n)$ random bits that guarantees the max-load of the* Uniform Greedy *scheme with d independent choices from $\mathcal{H}$ is $\log_d \log n + O\big(c + \frac{m}{n}\big)$ with probability at least $1 - n^{-c}$ for any m balls in U.*

Proof. We specify the parameters of $\mathcal{H}$ as follows: $k_g = 10c(\log_d \log m + \log_d(2+2c) + 5 + 3c), k = \log_2 \frac{\log n}{3 \log\log n}$, $\delta_2 = \log n^{-C \log n}$ for a large constant C, and $\delta_1 = 1/\text{poly}(n)$ such that Corollary 8.3.5 holds with probability at least $1 - n^{-c-1}$. Let $h^{(1)}, \ldots, h^{(d)}$ denote the d independent hash functions from $\mathcal{H}$ with the above parameters, where each

$$h^{(j)}(x) = \big(h_1^{(j)}(x) \circ h_2^{(j)}(x) \circ \cdots \circ h_k^{(j)}(x) \circ h_{k+1}^{(j)}(x)\big) \oplus g^{(j)}(x).$$

We use the notation g to denote $\{g^{(1)}, g^{(2)}, \ldots, g^{(d)}\}$ in the d choices and h_i to denote the group of hash functions $\{h_i^{(1)}, \ldots, h_i^{(d)}\}$ in this proof.

We bound the probability that any symmetric witness tree of height $l = \lceil \log_d \log m + \log_d(2+2c) + 5 + 3c \rceil$ with leaves of height at least $b = 10d \cdot \frac{m}{n} + 1$ exists in $h^{(1)}, \ldots, h^{(d)}$. Similar to the proof of Lemma 8.2.5, we bound the probability of pruned witness trees of height l in $h^{(1)}, \ldots, h^{(d)}$. We separate all pruned witness trees into two cases according to the number of edge collisions: pruned witness trees with at most $3c$ collisions and pruned witness trees with at least $3c$ collisions.

Pruned witness trees with at least $3c$ collisions. We start with a configuration C of pruned witness trees with height l and at least $3c$ collisions. Let $e_1, \ldots, e_{3c}$ be the first $3c$ collisions in the BFS of C. Let C' be the induced subgraph of C that only contains nodes

in these edges $e_1, \ldots, e_{3c}$ and their ancestors in C. Therefore any pruned witness tree T of configuration C exists in $h^{(1)}, \ldots, h^{(d)}$ only if the corresponding counterpart T' of T with configuration C' exists in $h^{(1)}, \ldots, h^{(d)}$. The existence of T' in $h^{(1)}, \ldots, h^{(d)}$ indicates that for every edge (u, v) in T', $h^{(1)}, \ldots, h^{(d)}$ satsify

$$h^{(i)}\big(T(u)\big) \in \big\{h^{(1)}\big(T(v)\big), \ldots, h^{(d)}\big(T(v)\big)\big\} \text{ when } v \text{ is the } i\text{th child of } u. \tag{8.4}$$

Notice that the number of edges in C' and T' is at most $3c \cdot 2l + 3c = 2l(3c+1) \leq k_g/2$.

Because $h^{(1)}, \ldots, h^{(d)}$ are k_g-wise independent, We bound the probability that all edges of T' satisfy (8.4) in $h^{(1)}, \ldots, h^{(d)}$ by

$$\prod_{(u,v)\in T'} (\frac{d}{n}) = (\frac{d}{n})^{|C'|+3c-1}.$$

Now we apply a union bound over all choices of balls in C'. There are at most $m^{|C'|}$ choices of balls in the nodes of C'. Therefore we bound the probability that any witness with at least $3c$ collisions survives in k_g-wise independent functions by

$$(\frac{d}{n})^{|C'|+3c-1} \cdot m^{|C'|} \leq (\frac{d}{n})^{3c-1} \cdot (\frac{m}{n} \cdot d)^{|C'|} \leq (\frac{d}{n})^{3c-1} \cdot (\frac{m}{n} \cdot d)^{3c\cdot(2l+1)} \leq n^{-2c}.$$

Next we apply a union bound over all configurations C'. Because there are at most $(1 + d + \cdots + d^l)^{2\cdot 3c} \leq n$ configurations of $3c$ collisions, with probability at least $1 - n^{-c}$, there is no pruned witness trees with at least $3c$ collision and height l exists in $h^{(1)}, \ldots, h^{(d)}$.

Pruned witness trees with at most $3c$ collisions. We fix a configuration C of pruned witness trees with height l and less than $3c$ collisions. Next we bound the probability that any pruned witness trees in this configuration C with leaves of height at least b exists in $h^{(1)}, \ldots, h^{(d)}$.

We extensively use the fact that after fixing g and $h_1 \circ \cdots \circ h_k$, at most $d(1.01 \log^3 n \cdot \frac{m}{n})$ elements in $h^{(1)}, \ldots, h^{(d)}$ are mapped to any bin of $[n/\log^3 n]$ from Corollary 8.3.5. Another

property is the number of leaves in C: because there are at most $3c$ collisions in C, C has at least $d^{l-3c} \in [d^5(2+2c)\log m, d^6(2+2c)\log m]$ leaves. On the other hand, the number of leaves is at least $\frac{|C|-3c}{2}$.

For a pruned witness tree T with configuration C, T exists in $h^{(1)}, \ldots, h^{(d)}$ only if

$$\forall (u,v) \in C, h^{(i)}\big(T(u)\big) \in \big\{h^{(1)}\big(T(v)\big), \ldots, h^{(d)}\big(T(v)\big)\big\} \text{ when } v \text{ is the } i\text{th child of } u. \quad (8.5)$$

We restate the above condition on the prefixes and suffixes of $h^{(1)}, \ldots, h^{(d)}$ separately. Let $g_p(x)$ denote the first $\log n - 3\log\log n$ bits of $g(x)$ and $g_s(x)$ denote the last $3\log\log n$ bits of $g(x)$, which matches $h_1(x) \circ \cdots \circ h_k(x)$ and $h_{k+1}(x)$ separately. Since $h^{(i)}(x) = \big(h_1^{(i)}(x) \circ \cdots \circ h_{k+1}^{(i)}(x)\big) \oplus g^{(i)}(x)$, property (8.5) indicates that the prefixes of the balls $b_u = T(u)$ and $b_v = T(v)$ satisfy

$$\big(h_1^{(i)}(b_u) \circ \cdots \circ h_k^{(i)}(b_u)\big) \oplus g_p^{(i)}(b_u) \in \left\{\big(h_1^{(i)}(b_v) \circ \cdots \circ h_k^{(i)}(b_v)\big) \oplus g_p^{(i)}(b_v)\big|_{i\in[d]}\right\}. \quad (8.6)$$

and their suffixes satisfy

$$h_{k+1}^{(i)}(b_u) \oplus g_s^{(i)}(b_u) \in \left\{h_{k+1}^{(1)}(b_v) \oplus g_s^{(1)}(b_v), \ldots, h_{k+1}^{(d)}(b_v) \oplus g_s^{(d)}(b_v)\right\}. \quad (8.7)$$

Let $\mathcal{F}_T$ be the subset of witness trees in the configuration C whose edges satisfy the condition (8.6) in preffixes $h_{(1)}, \ldots, h_{(k)}$, i.e.,

$$\mathcal{F}_T = \{T | \text{configuration}(T) = C \text{ and } (u,v) \text{ satisfies (8.6) } \forall (u,v) \in T\}.$$

We show that

$$|\mathcal{F}_T| \le m \cdot (d \cdot 1.01\log^3 n \cdot \frac{m}{n})^{|C|-1}.$$

The reason is as follows. There are m choices of balls for the root u in C. For the ith child v of the root u, we have to satisfy the condition (8.6) for (u,v). For a fixed bin $\big(h_1^{(i)}(b_u) \circ \cdots \circ h_k^{(i)}(b_u)\big) \oplus g_p^{(i)}(b_u)$, there are at most $1.01 \cdot \log^3 n \cdot \frac{m}{n}$ elements from each hash

function $h^{(j)}$ mapped to this bin from Corollary 8.3.5. Hence there are at most $d \cdot 1.01 \log^3 n \cdot \frac{m}{n}$ choices for each child of u. Then we repeat this arguments for all non-leaf nodes in C.

Next we consider the suffixes $h_{k+1}^{(1)}, \ldots, h_{k+1}^{(d)}$. We first calculate the probability that any possible witness tree in $\mathcal{F}_T$ survives in $h_{k+1}^{(1)}, \ldots, h_{k+1}^{(d)}$ from t-wise independence for $t = 5b \cdot d^{l+2} = O(\log n)$. After fixing g_s, for a possible witness tree T in $\mathcal{F}_T$, $h_{k+1}^{(1)}, \ldots, h_{k+1}^{(d)}$ satisfy (8.7) for every edge $(u, v) \in C$ with probability $\frac{d}{\log^3 n}$ in $t/2$-wise independent distributions because the number of edges in C is less than $t/2$.

For each leaf v in T, we bound the probability that its height is at least $b = 10d \cdot \frac{m}{n} + 1$ by $2^{-3d^2} \cdot (\frac{n}{m})^{2d}$ in $(b \cdot d + 1)$-wise independence. Given a choice $i \in [d]$ of leaf v, we fix the bin to be $h^{(i)}(v)$. Then we bound the probability that there are at least $b - 1$ balls $w_1, \ldots, w_{b-1}$ in this bin excluding all balls in the tree by

$$\sum_{w_1 : w_1 < v, w_1 \notin T} \sum_{w_2 : w_1 < w_2 < v, w_2 \notin T} \cdots \sum_{w_{b-1} : w_{b-2} < w_{b-1} < v, w_{b-1} \notin T} \Pr[h^{(i)}(v) = h^{(j_1)}(w_1) = \cdots = h^{(j_{b-1})}(w_{b-1})]$$
$$\leq \frac{\binom{1.01d \cdot \log^3 n \cdot \frac{m}{n}}{b-1}}{(\log^3 n)^{b-1}} \leq \frac{(1.01d \cdot \frac{m}{n})^{b-1}}{(b-1)!} \leq (\frac{3}{4})^{b-1}.$$

For all d choices of this leaf v, this probability is at most $(\frac{3}{4})^{(b-1) \cdot d} \leq 2^{-3d^2} \cdot (\frac{n}{m})^{2d}$.

Because $w_1, \ldots, w_b$ are not in the tree T for every leaf, they are disjoint and independent with the events of (8.7) in T, which are over all edges in the tree. Hence we could multiply these two probability together in t-wise independence given $t/2 \geq (b \cdot d + 1) \cdot$ number of leaves. Then we apply a union bound over all possible pruned witness trees in $\mathcal{F}_T$ to bound the probability (in the t-wise independence) that there is one witness tree of

height l whose leaves have height at least $10d \cdot \frac{m}{n} + 1$ by

$$\begin{aligned}
&|\mathcal{F}_T| \cdot (\frac{d}{\log^3 n})^{|C|-1} \cdot \left((\frac{3}{4})^{b\cdot d}\right)^{\text{number of leaves}} \\
\leq & m \left(1.01d \cdot \log^3 n \cdot \frac{m}{n} \cdot \frac{d}{\log^3 n}\right)^{|C|} \cdot \left(2^{-3d^2} \cdot (\frac{n}{m})^{2d}\right)^{\frac{|C|-3c}{2}} \\
\leq & m \cdot \left(2d^2 \cdot \frac{m}{n}\right)^{|C|} \cdot \left(2^{-3d^2} \cdot (\frac{n}{m})^{2d}\right)^{|C|/3} \\
\leq & m \cdot 2^{-|C|/3} \leq n^{-c-1}.
\end{aligned}$$

Finally we replace the t-wise independence by a δ_2-biased space for $\delta_2 = n^{-c-1} \cdot (\log^3 n)^{-t}/|\mathcal{F}_T| = (\log n)^{-O(\log n)}$. We apply Claim 8.3.6 to all possible pruned witness tress in $\mathcal{F}_T$: in δ_2-biased spaces, the probability of the existence of any height-l witness tree with leaves of height at least $b = 10d \cdot \frac{m}{n} + 1$ is at most

$$n^{-c-1} + |\mathcal{F}_T| \cdot \delta_2 \cdot (\log^3 n)^t \leq 2n^{-c-1}.$$

Then we apply a union bound on all possible configurations with at most $3c$ collisions:

$$(d^{l+1})^{|3c|} \cdot 2n^{-c-1} \leq 0.5n^{-c}.$$

From all discussion above, with probability at least $1 - n^{-c}$, there is no ball of height more than $l + b = \log_d \log n + O(1)$. □

8.5 The *Always-Go-Left* Scehme

We show that the hash family in Section 8.3 with proper parameters also achieves a max-load of $\frac{\log \log n}{d \log \phi_d} + O(1)$ in the *Always-Go-Left* scheme [Voc03] with d choices, where $\phi_d > 1$ is the constant satisfying $\phi_d^d = 1 + \phi_d + \cdots + \phi_d^{d-1}$. We define the *Always-Go-Left* scheme [Voc03] as follows:

Definition 8.5.1 (*Always-Go-Left* with d choices)**.** *Our algorithm partition the bins into d groups $G_1, \ldots, G_d$ of the same size n/d. Let $h^{(1)}, \ldots, h^{(d)}$ be d functions from U to $G_1, \ldots, G_d$ separately. For each ball b, the algorithm consider d bins $\{h^{(1)}(b) \in G_1, \ldots, h^{(d)}(b) \in G_d\}$ and chooses the bin with the least number of balls. If there are several bins with the least number of balls, our algorithm always choose the bin with the smallest group number.*

We define asymmetric witness trees for the *Always-Go-Left* mechanism such that a ball of height $l + C$ in the *Always-Go-Left* scheme indicates that there is an asymmetric witness tree of height l whose leaves have height at least C. For an asymmetric witness tree T, the height of T is still the *shortest* distance from the root to its leaves.

Definition 8.5.2 (Asymmetric Witness tree)**.** *The asymmetric witness tree T of height l in group G_i is a d-ary tree. The root has d children where the subtree of the jth child is an asymmetric witness tree in group G_j of height $(l - 1_{j \geq i})$.*

Given d functions $h^{(1)}, \ldots, h^{(d)}$ from U to $G_1, \ldots, G_d$ separately, a ball b with height more than $l + C$ in a bin of group G_i indicates an asymmetric witness tree T of height l in G_i whose leaves have height at least C. Each node of T corresponds to a ball, and the root of T corresponds to the ball b. A ball u in T has a ball v as its jth child iff when we insert the ball u in the Always-Go-Left *mechanism, v is the top ball in the bin $h^{(j)}(u)$. Hence $v < u$ and $h^{(j)}(u) = h^{(j)}(v)$ when the jth child of u is v.*

For an asymmetric witness tree T of height l in group G_i, We use the height l and the group index $i \in [d]$ to determine its size. Let $f(l, i)$ be the size of a *full* asymmetric witness tree of height l in group G_i. From the definition, we have $f(0, i) = 1$ and

$$f(l, i) = \sum_{j=1}^{i-1} f(l, j) + \sum_{j=i}^{d} f(l-1, j).$$

Let $g\big((l-1) \cdot d + i\big) = f(l, i)$ such that

$$g(n) = g(n-1) + g(n-2) + \cdots + g(n-d).$$

We know there exist $c_0 > 0$, $c_1 = O(1)$, and $\phi_d > 1$ satisfying

$$\phi_d^d = 1 + \phi_d + \cdots + \phi_d^{d-1} \text{ such that } g(n) \in [c_0 \cdot \phi_d^n, c_1 \cdot \phi_d^n].$$

Hence

$$f(l,i) = g\big((l-1)\cdot d + i\big) \in [c_0 \cdot \phi_d^{(l-1)d+i}, c_1 \cdot \phi_d^{(l-1)d+i}].$$

Similar to the pruned witness tree of a symmetric witness tree, we use the same process in Definition 8.2.3 to obtain the pruned asymmetric witness tree of an asymmetric witness tree.

Vöcking in [Voc03] showed that in a perfectly random hash function, the maximum load is $\frac{\log\log n}{d\log\phi_d} + O(1)$ with high probability given any n balls. We outline Vöcking's argument for *distinct balls* here: let b be a ball of height $l+4$ for $l = \frac{\log\log n + \log(1+c)}{d\log\phi_d} + 1$. Without loss of generality, we assume that b is in the first group G_1. By the definition of the asymmetric witness tree, there exists a tree T in G_1 with root b and height l whose leaves have height at least 4. For each ball u and its ith ball v, the hash function $h^{(i)}$ satisfies $h^{(i)}(u) = h^{(i)}(v)$. Similar to (8.2), we apply a union bound on all possible witness trees of height l in this configuration to bound the probability by

$$n^{f(l,1)} \cdot (\frac{d}{n})^{f(l,1)-1} \cdot (\frac{1}{3^d})^{\text{number of leaves in } f(l,1)},$$

which is less than n^{-c} given $f(l,1) = \Theta(\phi_d^{(l-1)d+1}) = \Theta\big((1+c)\log n\big)$.

We prove our derandomization of Vöcking's argument here.

Theorem 8.5.3. *For any $m = O(n)$, any constants $c > 1$ and $d \geq 2$, there exist a constant $\phi_d \in (1.61, 2)$ and a hash family $\mathcal{H}$ in Construction 8.3.1 with $O(\log n \log\log n)$ random bits such that for any m balls in U, with probability at least $1 - n^{-c}$, the max-load of the* Always-Go-Left *mechanism with d independent choices from $\mathcal{H}$ is $\frac{\log\log n}{d\log\phi_d} + O(c + \frac{m}{n})$.*

Proof. Let l be the smallest integer such that $c_0\phi_d^{ld} \geq 10(2+2c)\log m$ and $b = 10d \cdot \frac{m}{n} + 1$. We bound the probability of a witness tree of height $l + 3c + 1$ whose leaves have height

more than b in $h^{(1)}, \ldots, h^{(d)}$ during the *Always-Go-Left* scheme. Notice that there is a ball of height $l+b+3c+1$ in any bin of $G_2, G_3, \ldots, G_d$ indicates that there is a ball of the same height in G_1.

We choose the parameters of $\mathcal{H}$ as follows: $k_g = 20c{\cdot}d{\cdot}(l{+}b{+}1{+}3c) = O(\log\log n), k = \log_2(\log n/3\log\log n)$, $\delta_1 = 1/\text{poly}(n)$ such that Corollary 8.3.5 happens with probability at most n^{-c-1}, and the bias $\delta_2 = \log n^{-O(\log n)}$ of h_{k+1} later. We set h_{k+1} to be a hash function from U to $[\log^3 n/d]$ and g to be a function from U to $[n/d]$ such that

$$h^{(j)} = \left(h_1^{(j)} \circ h_1^{(j)} \circ \cdots \circ h_k^{(j)} \circ h_{k+1}^{(j)}\right) \oplus g^{(j)}$$

is a map from U to G_j of $[n/d]$ bins for each $j \in d$.

We use $h^{(1)}, \ldots, h^{(d)}$ to denote d independent hash functions from $\mathcal{H}$ with the above parameters. We use the notation of h_i to denote the group of hash functions $\{h_i^{(1)}, \ldots, h_i^{(d)}\}$ in this proof. We assume Corollary 8.3.5 and follow the same argument in the proof of Theorem 8.4.1. We bound the probability of witness trees from 2 cases depending on the number of collisions.

Pruned witness trees with at least $3c$ collisions: Given a configuration C with at least $3c$ collisions, we consider the first $3c$ collisions $e_1, \ldots, e_{3c}$ in the BFS of C. Let C' be the induced subgraph of C that only contains all vertices in $e_1, \ldots, e_{3c}$ and their ancestors in C. Therefore C survives under $h^{(1)}, \ldots, h^{(d)}$ only if C' survives under $h^{(1)}, \ldots, h^{(d)}$.

Observe that $|C'| \leq 3c \cdot 2 \cdot \left(d \cdot \text{height}(T)\right)$. There are at most $m^{|T'|}$ possible instantiations of balls in C'. For each instantiation T of C', because $k_g \geq 2 \cdot$ number of edges $= 2(|C'| + 3c - 1)$, we bound the probability that any instantiation of C' survives in h by

$$m^{|C'|} \cdot (\frac{d}{n})^{\text{number of edges}} = m^{|C'|} \cdot (\frac{d}{n})^{|C'|+3c-1} \leq (dm/n)^{|C'|} \cdot (\frac{d}{n})^{3c-1} \leq n^{-2c}.$$

At the same time, there are at most $(|T|^2)^{3c} = \text{poly}(\log n)$ configurations of C'. Hence we bound the probability of any witness with at least $3c$ collisions surviving by n^{-c}.

Pruned witness tree with less than $3c$ collisions: We fix a configuration C of witness tree in group G_1 with height $l+1+3c$ and less than $3c$ collisions. Thus $|C| \in [f(l+1,1), f(l+1+3c,1)]$.

Let $\mathcal{F}_T$ be the subset of possible asymmetric witness tree with configuration C after fixing the prefixes $h_1, h_2, \ldots, h_k$. For any $T \in \mathcal{F}_T$, each edge (u,v) has to satisfy $h^{(i)}\big(T(u)\big) = h^{(i)}\big(T(v)\big)$ in the *Always-Go-Left* scheme when v is the ith child of u. This indicates their prefixes are equal:

$$h_1^{(i)}\big(T(u)\big) \circ \cdots \circ h_k^{(i)}\big(T(u)\big) = h_1^{(i)}\big(T(v)\big) \circ \cdots \circ h_k^{(i)}\big(T(v)\big).$$

From the same argument in the proof of Theorem 8.4.1, we bound

$$|\mathcal{F}_T| \leq m \cdot (1.01 \log^3 n \cdot \frac{m}{n})^{|C|-1}$$

under $h_1, h_2, \ldots, h_k$ from Corollary 8.3.5.

We first consider h_{k+1} as a t-wise independent distribution from U to $[\log^3 n/d]$ for $t = 5bd \cdot f(l+3c+1,1) = O(\log m)$ then move to δ_2-biased spaces. For each asymmetric witness tree, every edge (u,v) maps to the same bin w.p. $d/\log^3 n$ in h_{k+1}.

For each leaf, its height is at least b if each bin in its choices has height at least $b-1$, which happens with probability at most

$$\left(\frac{\binom{1.01\cdot\log^3 n\cdot\frac{m}{n}}{b-1}}{(\log^3 n/d)^{b-1}}\right)^d \leq \left(\frac{(1.01d\cdot\frac{m}{n})^{b-1}}{(b-1)!}\right)^d \leq 2^{-3d^2}\cdot(\frac{n}{m})^{2d}$$

from the proof of Theorem 8.4.1.

Because these two types of events are on disjoint subsets of balls, the probability that any possible asymmetric witness tree in $\mathcal{F}_T$ exists in t-wise independent distributions over the suffixes is at most

$$\begin{aligned}|\mathcal{F}_T| \cdot \left(\frac{d}{\log^3 n}\right)^{|C|-1} \cdot \left(2^{-3d^2}\cdot(\frac{n}{m})^{2d}\right)^{\frac{(d-1)(|C|-3c)}{d}} &\leq m \cdot \left(1.01d\cdot\frac{m}{n}\right)^{|C|} \cdot \left(2^{-3d^2}\cdot(\frac{n}{m})^{2d}\right)^{|C|/3} \\ &\leq m \cdot 2^{-f(l+1,1)} \leq n^{-c-1}.\end{aligned}$$

We choose $\delta_2 = n^{-c-1} \cdot (\log^3 n/d)^{-t}/|\mathcal{F}_T| = (\log n)^{-O(\log n)}$ such that in δ_2-biased spaces, any possible asymmetric witness tree in $\mathcal{F}_T$ exists h_{k+1} is at most happens with probability at most $n^{-c-1} + |\mathcal{F}_T| \cdot \delta_2 \cdot (\log^3/d)^{bd \cdot f(l+3c+1,1)} \leq 2n^{-c-1}$. At the same time, the number of possible configurations is at most $(f(l+3c+1,1)^2)^{3c} \leq 0.1n$.

From all discussion above, with probability at most n^{-c}, there exists a ball in the *Always-Go-Left* mechanism with height at least $l + b + 3c + 1 = \frac{\log n \log n}{d \log \phi_d} + O(1)$. □

8.6 Heavy Load

We consider the derandomization of the 1-choice scheme when we have $m = n \cdot \text{poly}(\log n)$ balls and n bins. From the Chernoff bound, w.h.p, the max-load among n bins is $\frac{m}{n}\left(1 + O(\sqrt{\log n} \cdot \sqrt{\frac{n}{m}})\right)$ when we throw $m > n \log n$ balls into n bins independently at random. We modify the hash function from [CRSW13] with proper parameters for $m = \text{poly}(\log n) \cdot n$ balls and prove the max-load is still $\frac{m}{n}\left(1 + O(\sqrt{\log n} \cdot \sqrt{\frac{n}{m}})\right)$. We assume $m = \log^a n \cdot n$ for a constant $a \geq 1$ in the rest of this section.

Theorem 8.6.1. *For any constant $c > 0$ and $a \geq 1$, there exist a constant C and a hash function from U to $[n]$ generated by $O(\log n \log\log n)$ random bits such that for any $m = \log^a n \cdot n$ balls, with probability at least $1 - n^{-c}$, the max-load of the n bins in the 1-choice scheme with the hash function h is at most $\frac{m}{n}\left(1 + C \cdot \sqrt{\log n} \cdot \sqrt{\frac{n}{m}}\right)$.*

We omit g in this section and change $h_1, \ldots, h_{k+1}$ with different parameters. We choose $k = \log \frac{\log n}{(2a)\log\log n}$, h_i to denote a hash function from U to $[n^{2^{-i}}]$ for $i \in [k]$, and h_{k+1} to denote a hash function from U to $[n^{2^{-k}}] = [\log^{2a} n]$ such that $h_1 \circ h_2 \circ \cdots \circ h_k \circ h_{k+1}$ constitute a hash function from U to $[n]$. We set $\beta = 4(c+2)\sqrt{\log n}\sqrt{\frac{n}{m}}$. For convenience, we still think $h_1 \circ h_2 \circ \cdots \circ h_i$ as a hash function maps to $n^{1-2^{-i}}$ bins for any $i \leq k$. In this section, we still use δ_1-biased spaces on $h_1, \ldots, h_k$ and a δ_2-biased space on h_{k+1} for $\delta_1 = 1/\text{poly}(n)$ and $\delta_2 = (\log n)^{-O(\log n)}$.

Claim 8.6.2. *For any constant $c > 0$, there exists $\delta_1 = 1/\mathrm{poly}(n)$ such that given $m = \log^a n \cdot n$ balls, with probability $1 - n^{-c-1}$, for any $i \in [k]$ and any bin $b \in [n^{1-2^{-i}}]$, there are less than $\prod_{j \le i}(1 + \frac{\beta}{(k+2-i)^2}) \cdot \frac{m}{n} \cdot n^{2^{-i}}$ balls in this bin.*

Proof. We still use induction on i. The base case is $i = 0$. Because there are at most m balls, the hypothesis is true.

Suppose it is true for $i = l$. Now we fix a bin and assume there are $s = \prod_{j \le l}(1 + \frac{\beta}{(k+2-i)^2}) \cdot \frac{m}{n} n^{2^{-l}} \le (1+\beta)\frac{m}{n} n^{2^{-l}}$ balls in this bin from the induction hypothesis. h_{l+1} maps these s balls to $t = n^{2^{-(l+1)}}$ bins. We will prove that with high probability, every bin in these t bins of h_{l+1} contains at most $(1 + \frac{\beta}{(k+1-l)^2})s/t$ balls.

We use $X_i \in \{0, 1\}$ to denote whether ball i is in one fixed bin of $[t]$ or not. Hence $\Pr[X_i = 1] = 1/t$. Let $Y_i = X_i - \mathbb{E}[X_i]$. Therefore $\mathbb{E}[Y_i] = 0$ and $\mathbb{E}[|Y_i|^l] \le 1/t$ for any $l \ge 2$. Let $b = \beta 2^l$ for a large constant β later.

$$\begin{aligned}
\Pr_{D_{\delta_1}}[\sum_i X_i > (1 + \frac{\beta}{(k+1-l)^2})s/t] &\le \frac{\mathbb{E}_{D_{\delta_1}}[(\sum_i Y_i)^b]}{(\frac{\beta}{(k+1-l)^2}s/t)^b} \\
&\le \frac{\sum_{i_1,\ldots,i_b} \mathbb{E}_U[Y_{i_1} \cdots Y_{i_b}] + \delta_1 s^{2b}}{(\frac{\beta}{(k+1-l)^2}s/t)^b} \\
&\le \frac{2^b b!(s/t)^{b/2} + \delta_1 s^{2b}}{(\frac{\beta}{(k+1-l)^2}s/t)^b} \\
&\le \left(\frac{2b(s/t)}{(\frac{\beta}{(k+1-l)^2}s/t)^2}\right)^{b/2} + \delta_1 \cdot s^{2b}
\end{aligned}$$

We use these bounds $k = \log \frac{\log n}{(2l)\log\log n} < \log\log n$, $b < \beta 2^k < \frac{\beta \log n}{(2l)\log\log n}$ and $n^{2^{-l-1}} \ge n^{2^k} \ge$

$\log^{2l} n \geq (m/n)^2$ to simplify the above bound by

$$\begin{aligned}
&\left(\frac{2\log n}{\frac{\beta^2}{(\log\log n)^4}\cdot s/t}\right)^{b/2} + \delta_1 s^{2b} \\
\leq &\left(\frac{2\log^2 n}{(\log n\cdot\frac{n}{m})\cdot(\frac{m}{n}n^{2^{-l-1}})}\right)^{b/2} + \delta_1 s^{2b} \\
\leq &\left(\frac{1}{n^{0.5\cdot 2^{-l-1}}}\right)^{b/2} + \delta_1 s^{2b} \\
\leq &n^{-0.5\cdot 2^{-l-1}\cdot\beta 2^l/2} + \delta_1\left(\frac{2m}{n}n^{2^{-l}}\right)^{2\beta 2^l} \leq n^{-\beta/8} + \delta_1\cdot n^{6\beta}.
\end{aligned}$$

Hence we choose the two parameters $\beta > 8(c+2)$ and $\delta_1 = n^{-6\beta-c-2}$ such that the above probability is bounded by $2n^{-c-2}$. Finnally, we apply the union bound on i and all bins. □

Proof of Theorem 8.6.1. We first apply Claim 8.6.2 to $h_1, \ldots, h_k$.

In h_{k+1}, we first consider it as a $b = 16(c+2)^2\log n = O(\log n)$-wise independent distribution that maps $s < \prod_{j\leq k}(1+\frac{\beta}{(k+2-i)^2})\cdot\frac{m}{n}n^{2^{-k}}$ balls to $t = n^{2^{-k}}$ bins. From Lemma 8.1.6 and Theorem 5 (I) in [SSS95], we bound the probability that one bin receives more than $(1+\beta)s/t$ by $e^{\beta^2\cdot\mathbb{E}[s/t]/3} \leq n^{-c-2}$ given $b \geq \beta^2\,\mathbb{E}[s/t]$.

Then we choose $\delta_2 = (\log n)^{-b\cdot 5a} = (\log n)^{-O(\log n)}$ such that any δ_2-biased space from $[2\frac{m}{n}\log^{2a} n]$ to $[\log^{2a} n]$ is $\delta_2\cdot\binom{2\frac{m}{n}\log^{2a} n}{\leq b}\cdot(\log^{2a} n)^b < n^{-c-2}$-close to a b-wise independent distribution. Hence in h_{k+1}, with probability at most $2\cdot n^{-c-2}$, there is one bin that receives more than $(1+\beta)s/t$ balls. Overall, the number of balls in any bin of $[n]$ is at most

$$\prod_{i\leq k}(1+\frac{\beta}{(k+2-i)^2})(1+\beta)\frac{m}{n} \leq (1+\sum_{i\leq k+1}\frac{\beta}{(k+2-i)^2})\frac{m}{n} \leq (1+2\beta)\frac{m}{n}.$$

□

Chapter 9

Constraint Satisfaction Problems Above Average with Global Cardinality Constraints

In this chapter, we consider the constraint satisfaction problem on $\{-1,1\}^n$ under a global cardinality constraint. For generality, we allow different constraints using different predicates.

Definition 9.0.1. *An instance $\mathcal{I}$ of a constraint satisfaction problem of arity d consists of a set of variables $V = \{x_1, \cdots, x_n\}$ and a set of m constraints $C_1, \cdots, C_m$. Each constraint C_i consists of d variables $x_{i_1}, \cdots, x_{i_d}$ and a predicate $P_i \subseteq \{-1,1\}^d$. An assignment on $x_{i_1}, \cdots, x_{i_d}$ satisfies C_i if and only if $(x_{i_1}, \cdots, x_{i_d}) \in P_i$. The value $\mathsf{val}_{\mathcal{I}}(\alpha)$ of an assignment α is the number of constraints in $C_1, \cdots, C_m$ that are satisfied by α. The goal of the problem is to find an assignment with maximum possible value.*

An instance $\mathcal{I}$ of a constraint satisfaction problem with a global cardinality constraint consists of an instance $\mathcal{I}$ of a CSP and a global cardinality constraint $\sum_{i\in[n]} x_i = (1-2p)n$ specified by a parameter p. The goal of the problem is to find an assignment of maximum possible value complying with the global cardinality constraint $\sum_{i\in[n]} x_i = (1-2p)n$. We denote the value of the optimal assignment by

$$OPT = \max_{\alpha:\sum_i \alpha_i=(1-2p)n} \mathsf{val}_{\mathcal{I}}(\alpha).$$

The average value AVG of $\mathcal{I}$ is the expected value of an assignment chosen uniformly at random among all assignments complying the global cardinality constraint

$$AVG = \mathop{\mathbb{E}}_{\alpha:\sum_i \alpha_i=(1-2p)n} [\mathsf{val}_{\mathcal{I}}(\alpha)].$$

Given an instance $\mathcal{I}$ of a constraint satisfaction problem of arity d, we associate a degree-at-most d multilinear polynomial $f_{\mathcal{I}}$ with $\mathcal{I}$ such that $f_{\mathcal{I}}(\alpha) = \mathsf{val}_{\mathcal{I}}(\alpha)$ for any $\alpha \in \{\pm 1\}^n$.

$$f_{\mathcal{I}}(x) = \sum_{i \in [m]} \sum_{\sigma \in P_i} \frac{\prod_{j \in [d]} (1 + \sigma_j \cdot x_{i,j})}{2^d}.$$

Notice that given an instance $\mathcal{I}$ and a global cardinality constraint $\sum_{i \in n} x_i = (1-2p)n$, the expectation of $\mathcal{I}$ under the global cardinality constraint $AVG = \mathbb{E}_D[f_{\mathcal{I}}]$ is different than its expectation in the uniform distribution, even for CSPs of arity 2 in the bisection constraint.

Definition 9.0.2. *In the satisfiability above Average Problem, we are given an instance of a CSP of arity d, a global cardinality constraint $\sum_{i \in n} x_i = (1-2p)n$, and a parameter t. We need to decide whether $OPT \geq AVG + t$ or not.*

In this chapter, we show that it is fixed-parameter tractable. Namely, given a parameter t and an instance of a CSP problem of arity d under a global cardinality constraint $\sum_{i \in n} x_i = (1-2p)n$, we design an algorithm that either finds a kernel on $O(t^2)$ variables or certifies that $OPT \geq AVG + t$.

Theorem 9.0.3 (Informal version of Theorem 9.3.1 and Theorem 9.5.1)**.** *For any integer constant d and real constant $p_0 \in (0, 1/2]$, given a d-ary CSP with n variables and $m = n^{O(1)}$ constraints, a global cardinality constraint $\sum_{i=1}^{n} x_i = (1-2p)n$ such that $p \in [p_0, 1-p_0]$, and an integer parameter t, there is an algorithm that runs in time $(n^{O(1)} + 2^{O(t^2)})$ and decides whether there is an assignment complying with the cardinality constraint to satisfy at least $(AVG + t)$ constraints or not.*

One important ingredient in the proof of our main theorem is the $2 \to 4$ hypercontractivity of low-degree multilinear polynomials in a correlated probability space. Let D_p be the uniform distribution on all assignments to the n variables complying with the cardinality constraint $\sum_{i=1}^{n} x_i = (1-2p)n$. We show the following inequality.

Theorem 9.0.4 (Informal version of Corollary 9.3.8 and Corollary 9.4.2)**.** *For any degree d multilinear polynomial f on variables $x_1, x_2, \ldots, x_n$, we have*

$$\mathop{\mathbb{E}}_{D_p}[f^4] \leq \mathrm{poly}(d) \cdot C_p^d \cdot \mathop{\mathbb{E}}_{D_p}[f^2]^2,$$

where the constant $C_p = \mathrm{poly}(\frac{1}{1-p}, \frac{1}{p})$.

The ordinary $2 \to 4$ hypercontractive inequality (see Section 9.1.1 for details of the inequality) has wide applications in computer science, e.g., invariance principles [MOO10], a lower bound on the influence of variables on Boolean cube [KKL88], and an upper bound on the fourth moment of low degree functions [AGK+11, MMZ15] (see [O'D14] for a complete introduction and more applications with the reference therein). The inequality admits an elegant induction proof, which was first introduced in [MOO05]; and the proof was later extended to different settings (e.g. to the low-level sum-of-squares proof system [BBH+14], and to more general product distributions [MMZ15]). All the previous induction proofs, to the best of our knowledge, rely on the local independence of the variables (i.e. the independence among every constant-sized subset of random variables). In the $2 \to 4$ hypercontractive inequality we prove, however, every pair of the random variables is correlated.

Because of the lack of pair-wise independence, our induction proof (as well as the proof to the main theorem (Theorem 9.0.3)) crucially relies on the analysis of the eigenvalues of several $n^{O(d)} \times n^{O(d)}$ set-symmetric matrices. We will introduce more details about this analysis in the next subsection.

Related work. Recently, Gutin and Yeo [GY10] showed that it is possible to decide whether there is an assignment satisfying more than $\lceil m/2+t \rceil$ constraints in time $\left(2^{O(t^2)} + O(m)\right)$ for the MaxBisection problem with m constraints and n variables. The running time was later improved to $\left(2^{O(t)} + O(m)\right)$ by Mnich and Zenklusen [MZ12]. However, observe that in the MaxBisection problem, the trivial randomized algorithm satisfies $AVG =$

$\left(\frac{1}{2}+\frac{1}{2(n-1)}\right)m$ constraints in expectation. Therefore, when $m \gg n$, our problem MaxBisection above average asks more than what was proved in [GY10, MZ12]. For the MaxCut problem without any global cardinality constraint, Crowston et al. [CJM15] showed that optimizing above the Edwards-Erdős bound is fixed-parameter tractable, which is comparable to the bound in our work, while our algorithm outputs a solution strictly satisfying the global cardinality constraint $\sum_{i=1}^{n} x_i = (1-2p)n$.

Independently, Filmus and Mossel [FM16] provided a hypercontractive inequality over D_p based on the log-Sobolev inequality due to Lee and Yau [LY98]. They utilized the property that harmonic polynomials constitute an orthogonal basis in D_p. In this chapter, we use parity functions and their Fourier coefficients to analyze the eigenspaces of Var_{D_p} and prove the hypercontractivity in D_p. Parity functions do not constitute an orthogonal basis in D_p, e.g., the n variables are not independent under any global cardinality constraint $\sum_{i=1}^{n} x_i = (1-2p)n$. However, there is another important component in the proof of our main theorem – we need to prove the variance of a random solution is high if the optimal solution is much above average, where parity functions play an important role in this component.

Organization. We review several basic tools like Fourier analysis and Johnson scheme in Section 9.1. Then we analyze the eigenspaces of $E_{D_p}[f^2]$ and $\mathsf{Var}_{D_p}(f)$ in Section 9.2. Next we consider CSPs under the bisection constraint in Section 9.3. We prove the hypercontractivity Theorem 9.0.4 in Section 9.4. Finally we consider an arbitrary global constraint in Section 9.5.

9.1 Notation and Tools

In this chapter, we only consider $f : \{\pm 1\}^n \to \mathbb{R}$. Let U denote the uniform distribution on $\{\pm 1\}^n$ and U_p denote the biased product distribution on $\{\pm 1\}^n$ such that each bit equals to -1 with probability p and equals to 1 with probability $1-p$.

For a random variable X with standard deviation σ, it is known that the fourth moment is necessary and sufficient to guarantee that there exists $x \in \mathrm{supp}(X)$ greater than $\mathbb{E}[X] + \Omega(\sigma)$ from [Ber97, AGK+11, O'D14]. We state this result as follows.

Lemma 9.1.1. *Let X be a real random variable. Suppose that $\mathbb{E}[X] = 0, \mathbb{E}[X^2] = \sigma^2 > 0$, and $\mathbb{E}[X^4] < b\sigma^4$ for some $b > 0$. Then $\Pr[X \geq \sigma/(2\sqrt{b})] > 0$.*

In this chapter, we always use D to denote the uniform distribution on all assignments to the n variables complying with the bisection constraint $\sum_{i=1}^{n} x_i = 0$ and D_p to denote the uniform distribution on all assignments complying with the cardinality constraint $\sum_{i=1}^{n} x_i = (1-2p)n$.

9.1.1 Basics of Fourier Analysis of Boolean functions

We state several basic properties of the Fourier transform for Boolean functions those will be useful in this chapter. We first introduce the standard Fourier transform in $\{\pm 1\}^n$, which will be used in Section 9.3 and 9.5. We will also use the p-biased Fourier transform in several proofs especially for the $2 \to 4$ hypercontractive inequality under D_p in Section 9.1.2, Section 9.2, and Section 9.4.

For the uniform distribution U, we define the inner-product on a pair of functions $f, g : \{\pm 1\}^n \to \mathbb{R}$ by $\langle f, g \rangle = \mathbb{E}_{x \sim U}[f(x)g(x)]$. Hence $\chi_S(x) = \prod_{i \in S} x_i$ over all subsets $S \subseteq [n]$ constitute an orthonormal basis for the functions from $\{\pm 1\}^n$ to $\mathbb{R}$. We simplify the notation by writing χ_S instead of $\chi_S(x)$. Hence every Boolean function has a unique multilinear polynomial expression $f = \sum_{S \subseteq [n]} \hat{f}(S)\chi_S$, where $\hat{f}(S) = \langle f, \chi_S \rangle$ is the coefficient of χ_S in f. In particular, $\hat{f}(\emptyset) = \mathbb{E}_{x \sim U}[f(x)]$. An important fact about Fourier coefficients is Parseval's identity, i.e., $\sum_S \hat{f}(S)^2 = \mathbb{E}_{x \sim U}[f(x)^2]$, which indicates $\mathsf{Var}_U(f) = \sum_{S \neq \emptyset} \hat{f}(S)^2$.

Given any Boolean function f, we define its degree to be the largest size of S with non-zero Fourier coefficient $\hat{f}(S)$. In this chapter, we focus on the multilinear polynomials

f with degree-at-most d. We use the Fourier coefficients of weight i to denote all Fourier coefficients $\{\hat{f}(S)|S \in \binom{[n]}{i}\}$ of size i character functions. For a degree-at-most d polynomial f, we abuse the notation f to denote a vector in the linear space $\mathsf{span}\{\chi_S|S \in \binom{[n]}{\leq d}\}$, where each coordinate corresponds to a character function χ_S of a subset S.

We state the standard Bonami Lemma for Bernoulli ± 1 random variables [Bon70, O'D14], which is also known as the $2 \to 4$ hypercontractivity for low-degree multilinear polynomials.

Lemma 9.1.2. *Let $f : \{-1,1\}^n \to \mathbb{R}$ be a degree-at-most d multilinear polynomial. Let $X_1, \cdots, X_n$ be independent unbiased ± 1-Bernoulli variables. Then*

$$\mathbb{E}[f(X_1, \cdots, X_n)^4] \leq 9^d \cdot \mathbb{E}[f(X_1, \cdots, X_n)^2]^2.$$

For the p-biased distribution U_p, we define the inner product on pairs of function $f, g : \{\pm 1\}^n \to \mathbb{R}$ by $\langle f, g\rangle = \mathbb{E}_{x \sim U_p}[f(x)g(x)]$. Then we define $\phi_i(x) = \sqrt{\frac{p}{1-p}}1_{x_i=1} - \sqrt{\frac{1-p}{p}}1_{x_i=-1}$ and $\phi_S(x) = \prod_{i \in S} \phi_i(x)$. We abuse the notation by writing ϕ_S instead of $\phi_S(x)$. It is straightforward to verify $\mathbb{E}_{U_p}[\phi_i] = 0$ and $\mathbb{E}_{U_p}[\phi_i^2] = 1$. Notice that $\phi_S\phi_T \neq \phi_{S\Delta T}$ unlike $\chi_S\chi_S = \chi_{S\Delta T}$ for all x. However, $\langle \phi_S, \phi_T\rangle = 0$ for different S and T under U_p. Thus we have the biased Fourier expansion $f(x) = \sum_{S \subseteq [n]} \hat{f}(S)\phi_S(x)$, where $\hat{f}(S) = \langle f, \phi_S\rangle$ in U_p. We also have $\hat{f}(\emptyset) = \mathbb{E}_{U_p}[f]$ and Parseval's identity $\sum_S \hat{f}(S)^2 = \mathbb{E}_{U_p}[f^2]$, which demonstrates $\mathsf{Var}_{U_p}(f) = \sum_{S \neq \emptyset} \hat{f}(S)^2$. We state two facts of ϕ_i that will be useful in the later section.

1. $x_i = 2\sqrt{p(1-p)} \cdot \phi_i + 1 - 2p$. Hence $\sum_i \phi_i(x) = 0$ for any x satisfying $\sum_i x_i = (1-2p)n$.
2. $\phi_i^2 = q \cdot \phi_i + 1$ for $q = \frac{2p-1}{\sqrt{p(1-p)}}$. Thus we write f as a multilinear polynomial of ϕ_i.

Observe that the largest size of $|T|$ with non-zero Fourier coefficient $\hat{f}(T)$ in the basis $\{\phi_S|S \in \binom{[n]}{\leq d}\}$ is equivalent to the degree of f defined in $\{\chi_S|S \in \binom{[n]}{\leq d}\}$. Hence we still

define the degree of f to be $\max_{S:\hat{f}(S)\neq 0}|S|$. We abuse the notation f to denote a vector in the linear space $\mathsf{span}\{\phi_S | S \in \binom{[n]}{\leq d}\}$.

For the biased distribution U_p, we know $\mathbb{E}_{U_p}[\phi_i^4] = \frac{p^2}{1-p} + \frac{(1-p)^2}{p} \geq 1$. Therefore we state the $2 \to 4$ hypercontractivity in the biased distribution U_p as follows.

Lemma 9.1.3. *Let $f : \{-1,1\}^n \to \mathbb{R}$ be a degree-at-most d multilinear polynomial of $\phi_1, \cdots, \phi_n$. Then*

$$\mathop{\mathbb{E}}_{U_p}[f(X_1, \cdots, X_n)^4] \leq \left(9 \cdot \frac{p^2}{1-p} + 9 \cdot \frac{(1-p)^2}{p}\right)^d \cdot \mathop{\mathbb{E}}_{U_p}[f(X_1, \cdots, X_n)^2]^2.$$

At last, we notice that the definition of ϕ_S is consistent with the definition of χ_S when $p = 1/2$. When the distribution U_p is fixed and clear, we use $\|f\|_2 = E_{x\sim U_p}[f(x)^2]^{1/2}$ to denote the L_2 norm of a Boolean function f. From Parseval's identity, $\|f\|_2$ is also $(\sum_S \hat{f}(S)^2)^{1/2}$. From the Cauchy-Schwarz inequality, one useful property is $\|fg\|_2 \leq \|f^2\|_2^{1/2}\|g^2\|_2^{1/2}$.

9.1.2 Distributions conditioned on global cardinality constraints

We will study the expectation and the variance of a low-degree multilinear polynomial f in D_p. Because ϕ_S is consistent with χ_S when $p = 1/2$, we fix the basis to be ϕ_S of the p-biased Fourier transform. We treat $q = \frac{2p-1}{\sqrt{p(1-p)}}$ as a constant and hide it in the big-Oh notation.

We first discuss the expectation of f under D_p. Because $\mathbb{E}_{D_p}[\phi_S]$ is not necessary 0 for any non-empty subset S, $\mathbb{E}_{D_p}[f] \neq \hat{f}(\emptyset)$. Let $\delta_S = \mathbb{E}_{D_p}[\phi_S]$. From symmetry, $\delta_S = \delta_{S'}$ for any S and S' with the same size. For convenience, we use $\delta_k = \delta_S$ for any $S \in \binom{[n]}{k}$. From the definition of δ, we have $\mathbb{E}_{D_p}[f] = \sum_S \hat{f}(S) \cdot \delta_S$.

For $p = 1/2$ and D, $\delta_k = 0$ for all odd k and $\delta_k = (-1)^{k/2}\frac{(k-1)!!}{(n-1)\cdot(n-3)\cdots(n-k+1)}$ for even k. We calculate it this way: pick any $T \in \binom{[n]}{k}$ and consider $\mathbb{E}_D[(\sum_i x_i)\chi_T] = 0$. This indicates

$$k \cdot \delta_{k-1} + (n-k)\delta_{k+1} = 0.$$

From $\delta_0 = 1$ and $\delta_1 = 0$, we could obtain δ_k for every $k > 1$.

For $p \neq 1/2$ and D_p under the global cardinality constraint $\sum_{i\in n} x_i = (1-2p)n$, we consider $\mathbb{E}_{D_p}[\phi_S]$, because $\sum_{i\in n} x_i = (1-2p)n$ indicates $\sum_i \phi_i = 0$. Thus we use $\delta_S = \mathbb{E}_{D_p}[\phi_S]$ and calculate it as follows: pick any $T \in \binom{[n]}{k}$ and consider $\mathbb{E}_{D_p}[(\sum_i \phi_i)\phi_T] = 0$. $\phi_i\phi_T = \phi_{T\cup i}$ for $i \notin T$; and $\phi_i\phi_T = q \cdot \phi_T + \phi_{T\setminus i}$ for $i \in T$ from the fact $\phi_i^2 = q \cdot \phi_i + 1$. We have

$$k \cdot \delta_{k-1} + k \cdot q \cdot \delta_k + (n-k)\delta_{k+1} = 0 \tag{9.1}$$

Remark 9.1.4. *For $p = 1/2$ and the bisection constraint, $q = 0$ and the recurrence relation becomes $k \cdot \delta_{k-1} + (n-k)\delta_{k+1} = 0$, which is consistent with the above characterization. Thus we abuse the notation δ_k when U_p is fixed and clear.*

From $\delta_0 = 1, \delta_1 = 0$, and the relation above, we can determine δ_k for every k. For example, $\delta_2 = -\frac{1}{n-1}$ and $\delta_3 = -\frac{\delta_2}{n-2} \cdot 2 \cdot q = \frac{2q}{(n-1)(n-2)}$. We bound δ_i as follows:

Claim 9.1.5. *For any $i \geq 1$, $\delta_{2i-1} = (-1)^i O(n^{-i})$ and $\delta_{2i} = (-1)^i \frac{(2i-1)!!}{n^i} + O(n^{-i-1})$.*

Proof. We use induction on i. Base Case: $\delta_0 = 1$ and $\delta_1 = 0$.

Because $\delta_{2i-2} = (-1)^{i-1}\Theta(n^{-i+1})$ and $\delta_{2i-1} = (-1)^i O(n^{-i})$, the major term of δ_{2i} is determined by δ_{2i-2}. We choose $k = 2i-1$ in the equation (9.1) to obtain

$$\delta_{2i} = (-1)^i \frac{(2i-1)!!}{(n-1)(n-3)\cdots(n-2i+1)} + O(\frac{1}{n^{i+1}}) = (-1)^i \frac{(2i-1)!!}{n^i} + O(\frac{1}{n^{i+1}}).$$

At the same time, from δ_{2i} and δ_{2i-1},

$$\delta_{2i+1} = (-1)^{i+1} q \cdot \frac{(2i)(2i-1)!! + (2i)(2i-2)(2i-3)!! + \cdots + (2i)!!}{n^{i+1}} + O(\frac{1}{n^{i+2}}) = (-1)^{i+1} O(\frac{1}{n^{i+2}}).$$

□

Now we turn to $\mathbb{E}_{D_p}[f^2]$ and $\mathsf{Var}_{D_p}[f]$ for a degree-at-most-d multilinear polynomial f. From the definition and the Fourier transform $f = \sum_S \hat{f}(S)\phi_S$,

$$\mathop{\mathbb{E}}_{D_p}[f^2] = \sum_{S,T} \hat{f}(S)\hat{f}(T)\delta_{S\Delta T}, \quad \mathsf{Var}_{D_p}(f) = \mathop{\mathbb{E}}_{D_p}[f^2] - \mathop{\mathbb{E}}_{D_p}[f]^2 = \sum_{S,T} \hat{f}(S)\hat{f}(T)(\delta_{S\Delta T} - \delta_S\delta_T).$$

We associate a $\binom{n}{\leq d} \times \binom{n}{\leq d}$ matrix A with $\mathbb{E}_{D_p}[f^2]$ that $A(S,T) = \delta_{S\Delta T}$. Hence $\mathbb{E}_{D_p}[f^2] = f^T \cdot A \cdot f$ from the definition when we think f is a vector in the linear space of $\mathsf{span}\{\phi_S | S \in \binom{[n]}{\leq d}\}$.

Similarly, we associate a $\binom{n}{\leq d} \times \binom{n}{\leq d}$ matrix B with $\mathsf{Var}_{D_p}(f)$ that $B(S,T) = \delta_{S\Delta T} - \delta_S \cdot \delta_T$. Hence $\mathsf{Var}_{D_p}(f) = f^T \cdot B \cdot f$. Notice that an entry (S,T) in A and B only depends on the size of S, T, and $S \cap T$.

Remark 9.1.6. *Because $B(\emptyset, S) = B(S, \emptyset) = 0$ for any S and $\mathsf{Var}_{D_p}(f)$ is independent with $\hat{f}(\emptyset)$, we could neglect $\hat{f}(\emptyset)$ in B such that B is a $(\binom{[n]}{d} + \cdots \binom{[n]}{1}) \times (\binom{[n]}{d} + \cdots \binom{[n]}{1})$ matrix. $\hat{f}(\emptyset)$ is the only difference between the analysis of eigenvalues in A and B. Actually, the difference $\delta_S \cdot \delta_T$ between $A(S,T)$ and $B(S,T)$ will not effect the analysis of their eigenvalues except the eigenvalue induced by $\hat{f}(\emptyset)$.*

In Section 9.2, we study the eigenvalues of $\mathbb{E}_{D_p}[f^2]$ and $\mathsf{Var}_{D_p}(f)$ in the linear space $\mathsf{span}\{\phi_S | S \in \binom{[n]}{\leq d}\}$, i.e., the eigenvalues of A and B.

9.1.3 Eigenspaces in the Johnson Schemes

We shall use a few characterizations about the eigenspaces of the Johnson scheme to analyze the eigenspaces and eigenvalues of A and B in Section 9.2 (please see [God10] for a complete introduction).

We divide A into $(d+1) \times (d+1)$ submatrices where $A_{i,j}$ is the matrix of $A(S,T)$ over all $S \in \binom{[n]}{i}$ and $T \in \binom{[n]}{j}$. For each diagonal matrix $A_{i,i}$, observe that $A_{i,i}(S,T)$ only depends on $|S \cap T|$ because of $|S| = |T| = i$, which indicates $A_{i,i}$ is in the association schemes, in particular, Johnson scheme.

Definition 9.1.7. *A matrix $M \in \mathbb{R}^{\binom{[n]}{r}\times\binom{[n]}{r}}$ is set-symmetric if for every $S,T \in \binom{[n]}{r}$, $M(S,T)$ depends only on the size of $|S \cap T|$.*

For $n, r \leq n/2$, let $\mathcal{J}_r \subseteq \mathbb{R}^{\binom{[n]}{r}\times\binom{[n]}{r}}$ be the subspace of all set-symmetric matrices. $\mathcal{J}_r$ is called the Johnson scheme.

Let r be a fixed integer and $M \in \mathbb{R}^{\binom{[n]}{r}\times\binom{[n]}{r}}$ be a matrix in the Johnson scheme $\mathcal{J}_r$. We treat a vector in $\mathbb{R}^{\binom{[n]}{r}}$ as a homogeneous degree r polynomial $f = \sum_{T\in\binom{[n]}{r}} \hat{f}(T)\phi_T$, where each coordinate corresponds to a r-subset. Although the eigenvalues of M depend on the entries of M, the eigenspaces of M are independent with M as long as M is in the Johnson scheme.

Fact 9.1.8. *There are $r+1$ eigenspaces $V_0, V_1, \cdots, V_r$ in M. For $i \in [r]$, the dimension of V_i is $\binom{n}{i} - \binom{n}{i-1}$; and the dimension of V_0 is 1. We define V_i through $\hat{f}(S)$ over all $S \in \binom{[n]}{i}$, although M and f only depend on $\{\hat{f}(T)|T \in \binom{[n]}{r}\}$. V_i is the linear space spanned by $\{\hat{f}(S)\phi_S|S \in \binom{[n]}{i}\}$ with the following two properties:*

1. *For any $T' \in \binom{[n]}{i-1}$, $\{\hat{f}(S)|S \in \binom{[n]}{i}\}$ satisfies that $\sum_{j\notin T'} \hat{f}(T' \cup j) = 0$ (neglect this property for V_0).*
2. *For any $T \in \binom{[n]}{r}$, $\hat{f}(T) = \sum_{S\in\binom{T}{i}} \hat{f}(S)$.*

It is straightforward to verify that the dimension of V_i is $\binom{n}{i} - \binom{n}{i-1}$ and V_i is an eigenspace in M. Notice that the homogeneous degree i polynomial $\sum_{S\in\binom{[n]}{i}} \hat{f}(S)\phi_S$ is an eigenvector of matrices in $\mathcal{J}_i$.

To show the orthogonality between V_i and V_j, it is enough to prove that

Claim 9.1.9. *For any $j \leq r$ and any $S \in \binom{[n]}{<j}$, $\sum_{T\in\binom{[n]}{j}:S\subset T} \hat{f}(T) = 0$ for any $f \in V_j$.*

Proof. We use induction on the size of S to show it is true.

Base Case $|S| = j - 1$: from the definition of f, $\sum_{T:S\subset T} \hat{f}(T) = 0$.

Suppose $\sum_{T:S\subset T} \hat{f}(T) = 0$ for any $S \in \binom{[n]}{k+1}$. We prove it is true for any $S \in \binom{[n]}{k}$:

$$\sum_{T:S\subset T} \hat{f}(T) = \frac{1}{j-|S|} \sum_{i\notin S} \sum_{T:(S\cup i)\subset T} \hat{f}(T) = 0.$$

□

9.2 Eigenspaces and Eigenvalues of $\mathbb{E}_{D_p}[f^2]$ and $\mathsf{Var}_{D_p}(f)$

In this section we analyze the eigenvalues and eigenspaces of A and B, following the approach of Grigoriev [Gri01b].

We fix any $p \in (0,1)$ with the global cardinality constraint $\sum_i x_i = (1-2p)n$ and use the p-biased Fourier transform in this section, i.e., $\{\phi_S | S \in \binom{[n]}{\leq d}\}$. Because χ_S is consistent with ϕ_S for $p = 1/2$, it is enough to study the eigenspaces of A and B in $\mathsf{span}\{\phi_S | S \in \binom{[n]}{\leq d}\}$. Since A can be divided into $(d+1)\times(d+1)$ submatrices where we know the eigenspaces of the diagonal submatrices from the Johnson scheme, we study the eigenspaces of A through the global cardinality constraint $\sum_i \phi_i = 0$ and the relations between eigenspaces of these diagonal matrices characterized in Section 9.1.3. We will focus on the analysis of A in most time and discuss about B in the end of this section.

We first show the eigenspace V'_{null} with an eigenvalue 0 in A, i.e., the null space of A. Because $\sum_i x_i = (1-2p)n$ in the support of D_p, $\sum_i \phi_i(x) = 0$ for any x in the support of D_p. Thus $(\sum_i \phi_i)h = 0$ for all polynomial h of degree-at-most $d-1$, which is in the linear subspace $\mathsf{span}\{(\sum_i \phi_i)\phi_S | S \in \binom{[n]}{\leq d-1}\}$. This linear space is the eigenspace of A with an eigenvalue 0; and its dimension is $\binom{n}{\leq d-1} = \binom{[n]}{d-1} + \binom{[n]}{d-2} + \cdots \binom{[n]}{0}$. By the same reason, V'_{null} is the eigenspace in B with an eigenvalue 0.

Let V_d be the largest eigenspace in $A_{d,d}$ on $\binom{[n]}{d} \times \binom{[n]}{d}$. We demonstrate how to find an eigenspace of A based on V_d. From the definition of V_d, for any $f_d \in V_d$, f_d satisfies that $\sum_{j \notin T} \hat{f}_d(T \cup j) = 0$ for any $T \in \binom{[n]}{d-1}$ from the property of the Johnson scheme. Thus, from Claim 9.1.9 and the fact that $A(S,T)$ only depends on $|S \cap T|$ given $S \in \binom{[n]}{i}$ and $T \in \binom{[n]}{d}$, we know $A_{i,d} f_d = \vec{0}$ for all $i \leq d-1$. We construct an eigenvector f in A from f_d as follows: $\hat{f}(S) = 0$ for all $S \in \binom{[n]}{<d}$ and $\hat{f}(T) = \hat{f}_d(T)$ for all $T \in \binom{[n]}{d}$, i.e., $f = (\vec{0}, f_d)$. It is straightforward to verify that $A(\vec{0}, f_d) = \lambda_d(\vec{0}, f_d)$, where the eigenvalue λ_d is the eigenvalue of V_d in $A_{d,d}$.

Then we move to V_{d-1} in $A_{d,d}$ and illustrate how to use an eigenvector in V_{d-1} to construct an eigenvector of A. For any $f_d \in V_{d-1}$, let $f_{d-1} = \sum_{S \in \binom{[n]}{d-1}} \hat{f}_{d-1}(S) \phi_S$ be the homogeneous degree $d-1$ polynomial such that $f_d = \sum_{T \in \binom{[n]}{d}} \left(\sum_{S \in \binom{T}{d-1}} \hat{f}_{d-1}(S) \right) \phi_T$. From Claim 9.1.9, $A_{i,d} f_d = 0$ for all $i < d-1$ and $A_{i,d-1} f_{d-1} = 0$ for all $i < d-2$. Observe that f_{d-1} is an eigenvector of $A_{d-1,d-1}$, although it is possible that the eigenvalue of f_{d-1} in $A_{d-1,d-1}$ is different than the eigenvalue of f_d in $A_{d,d}$. At the same time, from the symmetry of A and the relationship between f_d and f_{d-1}, $A_{d-1,d} f_d = \beta_0 f_{d-1}$ and $A_{d,d-1} f_{d-1} = \beta_1 f_d$ for some constants β_0 and β_1 only depending on δ and d. Thus we can find a constant $\alpha_{d-1,d}$ such that $(\vec{0}, f_{d-1}, \alpha_{d-1,d} f_d)$ becomes an eigenvector of A.

More directly, we determine the constant $\alpha_{d-1,d}$ from the orthogonality between $(\vec{0}, f_{d-1}, \alpha_{d-1,d} \cdot f_d)$ and the null space $\mathsf{span}\{(\sum_i \phi_i)\phi_S | S \in \binom{[n]}{\leq d-1}\}$. We pick any $T \in \binom{[n]}{d-1}$ and rewrite $(\sum_i \phi_i)\phi_T = \sum_{j \in T} \phi_{T \setminus j} + q \cdot |T| \cdot \phi_T + \sum_{j \notin T} \phi_{T \cup j}$. From the orthogonality,

$$
\begin{aligned}
& q|T| \cdot \hat{f}_{d-1}(T) + \sum_{j \notin T} \alpha_{d-1,d} \Big(\sum_{T' \in \binom{T \cup j}{d-1}} \hat{f}_{d-1}(T') \Big) = 0 \\
& \qquad \Rightarrow \left(q|T| + (n - |T|)\alpha_{d-1,d} \right) \hat{f}_{d-1}(T) + \alpha_{d-1,d} \left(\sum_{T'' \in \binom{T}{d-2}} \sum_{j \notin T} \hat{f}_{d-1}(T'' \cup j) \right) = 0.
\end{aligned}
$$

From the property of f_{d-1} that $\sum_{j\notin T''}\hat{f}_{d-1}(T''\cup j)=0$ for all $T''\in\binom{[n]}{d-2}$, we simplify it to

$$\big(q|T|+(n-2|T|)\alpha_{d-1,d}\big)\hat{f}_{d-1}(T)=0,$$

which determines $\alpha_{d-1,d}=\frac{-(d-1)q}{n-2d+2}$ directly.

Following this approach, we figure out all eigenspaces of A from the eigenspaces $V_0,V_1,\cdots,V_d$ in $A_{d,d}$. For convenience, we use V_k' for $k\le d$ to denote the kth eigenspace in A extended by V_k in $A_{d,d}$. We first choose the coefficients in the combination. Let $\alpha_{k,i}=0$ for all $i<k$, $\alpha_{k,k}=1$, and $\alpha_{k,k+1},\cdots,\alpha_{k,d}$ satisfy the recurrence relation (we will show the choices of α later):

$$i\cdot\alpha_{k,k+i-1}+(k+i)\cdot q\cdot\alpha_{k,k+i}+(n-2k-i)\alpha_{k,k+i+1}=0. \tag{9.2}$$

Then for every $f\in V_k$, the coefficients of $\hat{f}(T)$ over all $T\in\binom{[n]}{\le d}$ spanned by $\{\hat{f}(S)|S\in\binom{[n]}{k}\}$ satisfy the following three properties:

1. $\forall T\in\binom{[n]}{k-1},\sum_{j\notin T}\hat{f}(T\cup j)=0$ (neglect this property for V_0');
2. $\forall T\in\binom{[n]}{>k},\hat{f}(T)=\alpha_{k,|T|}\cdot\sum_{S\in\binom{T}{k}}\hat{f}(S)$;
3. $\forall T\in\binom{[n]}{<k},\hat{f}(T)=0$.

Now we show the recurrence relation of $\alpha_{k,k+i}$ from the fact that f is orthogonal to the null space of A. We consider $(\sum_i\phi_i)\phi_T$ in the null space for a subset T of size $k+i<d$ and simplify $(\sum_i\phi_i)\phi_T$ to $\sum_{j\in T}\phi_{T\setminus j}+q\cdot|T|\cdot\phi_T+\sum_{j\notin T}\phi_{T\cup j}$. We have

$$\sum_{j\in T}\alpha_{k,k+i-1}\sum_{S\in\binom{T\setminus j}{k}}\hat{f}(S)+(k+i)\cdot q\cdot\alpha_{k,k+i}\sum_{S\in\binom{T}{k}}\hat{f}(S)+\sum_{j\notin T}\alpha_{k,k+i+1}\sum_{S\in\binom{T\cup j}{k}}\hat{f}(S)=0\Rightarrow$$
$$\sum_{S\in\binom{T}{k}}\big(i\cdot\alpha_{k,k+i-1}+(k+i)q\cdot\alpha_{k,k+i}+(n-k-i)\alpha_{k,k+i+1}\big)\hat{f}(S)+\sum_{T'\in\binom{T}{k-1}}\alpha_{k,k+i+1}\sum_{j\notin T}\hat{f}(T'\cup j)=0.$$

Using the first property $\forall T' \in \binom{[n]}{k-1}, \sum_{j \notin T'} \hat{f}(T' \cup j) = 0$ to eliminate all S' not in T, We obtain

$$\left(i \cdot \alpha_{k,k+i-1} + (k+i) \cdot q \cdot \alpha_{k,k+i} + (n-2k-i)\alpha_{k,k+i+1}\right) \sum_{S \in \binom{T}{k}} \hat{f}(S) = 0.$$

Because $\sum_{S \in \binom{T}{k}} \hat{f}(S)$ is not necessary equal to 0 to satisfy the first property (actually $\sum_{S \in \binom{T}{k}} \hat{f}(S) = 0$ for all $T \in \binom{[n]}{k+i}$ indicates $\hat{f}(S) = 0$ for all $S \in \binom{[n]}{k}$), the coefficient is 0, which provides the recurrence relation in (9.2).

The dimension of V_k' is $\binom{[n]}{k} - \binom{[n]}{k-1}$ from the first property (It is straightforward to verify $\sum_{k=0}^{d} dim(V_k') + dim(V_{null}') = \sum_{k=0}^{d}(\binom{[n]}{k} - \binom{[n]}{k-1}) + \binom{[n]}{d-1} + \binom{[n]}{d-2} + \cdots \binom{[n]}{0} = \binom{[n]}{\leq d}$). The orthogonality between V_i' and V_j' follows from Claim 9.1.9 and the orthogonality of V_i and V_j.

Remark 9.2.1. *$V_1', \cdots, V_d'$ are the non-zero eigenspaces in B except for V_0'. For $f \in V_0'$, observe that $\hat{f}(T)$ only depends on the size of T and $\hat{f}(\emptyset)$. Hence for any polynomial $f \in V_0'$, f is a constant function over the support of D_p, i.e., $\mathsf{Var}_{D_p}(V_0') = 0$. Therefore V_0' is in the null space of B.*

We use induction on i to bound $\alpha_{k,k+i}$. From $\alpha_{k,k} = 1$ and the recurrence relation (9.2), the first few terms would be $\alpha_{k,k+1} = -\frac{kq}{n-2k}$ and $\alpha_{k,k+2} = -\frac{1+(k+1)q \cdot \alpha_{k,k+1}}{n-2k-1} = -\frac{1}{n-2k-1} + O(n^{-2})$.

Claim 9.2.2. *$\alpha_{k,k+2i} = (-1)^i \frac{(2i-1)!!}{n^i} + O(n^{-i-1})$ and $\alpha_{k,k+2i+1} = (-1)^{i+1} O(n^{-i-1})$.*

Proof. We use induction on i again. Base Case: $\alpha_{k,k} = 1$ and $\alpha_{k,k+1} = -\frac{kq}{n-2k}$.

From the induction hypothesis $\alpha_{k,k+2i-2} = (-1)^{i-1}\Theta(n^{-i+1})$ and $\alpha_{k,k+2i-1} = (-1)^i n^{-i}$, the major term of $\alpha_{k,k+2i}$ is determined $\alpha_{k,k+2i-2}$ such that $\alpha_{k,k+2i} = (-1)^i \frac{(2i-1)!!}{n^i} + O(n^{-i-1})$. Similarly, $\alpha_{k,k+2i+1} = (-1)^{i+1} O(n^{-i-1})$. □

Now we bound the eigenvalue of V_k'. For convenience, we think $0! = 1$ and $(-1)!! = 1$.

Theorem 9.2.3. *For any $k \in \{0, \cdots, d\}$, the eigenvalue of V_k' in A is $\sum_{even\ i=0}^{d-k} \frac{(i-1)!!(i-1)!!}{i!} \pm O(n^{-1})$. For any $k \in \{1, \cdots, d\}$, the eigenvalue of V_k' in B is the same $\sum_{even\ i=0}^{d-k} \frac{(i-1)!!(i-1)!!}{i!} \pm O(n^{-1})$.*

Proof. We fix a polynomial $f \in V_k'$ and $S \in \binom{[n]}{k}$ to calculate $\sum_{T \in \binom{[n]}{\leq d}} A(S,T)\hat{f}(T)$ for the eigenvalue of V_k' in A. From the fact $\hat{f}(T) = \alpha_{k,|T|} \cdot \sum_{S' \in \binom{T}{k}} \hat{f}(S')$, we expand $\sum_T A(S,T)\hat{f}(T)$ into the summation of $\hat{f}(S')$ over all $S' \in \binom{[n]}{k}$ with coefficients. From the symmetry of A, the coefficients of $\hat{f}(S')$ in the expansion only depends on the size of $S \cap S'$ (the sizes of S and S' are k). Hence we use τ_i to denote the coefficients of $\hat{f}(S')$ given $|S'\Delta S| = i$. Thus $\sum_T A(S,T)\hat{f}(T) = \sum_{S' \in \binom{[n]}{k}} \tau_{|S'\Delta S|}\hat{f}(S')$.

We calculate $\tau_0, \cdots, \tau_{2d}$ as follows. Because $|S'| = |S| = k$, $|S\Delta S'|$ is always even. For τ_0, we only consider T containing S and use $k+i$ to denote the size of T.

$$\tau_0 = \sum_{i=0}^{d-k} \binom{n-k}{i} \cdot \alpha_{k,k+i} \cdot \delta_i. \tag{9.3}$$

For τ_{2l}, we fix a subset S' with $S\Delta S' = 2l$ and only consider T containing S'. We use $k+i$ to denote the size of T and t to denote the size of the intersection of T and $S \setminus S'$.

$$\tau_{2l} = \sum_{i=0}^{d-k} \alpha_{k,k+i} \sum_{t=0}^{i} \binom{l}{t}\binom{n-k-2l}{i-t}\delta_{2l+i-2t}. \tag{9.4}$$

We will prove that $\tau_0 = \Theta(1)$ and $\tau_{2l} = O(n^{-l})$ for all $l \geq 1$ then eliminate all $S' \neq S$ in $\sum_T A(S,T)\hat{f}(T) = \sum_{S' \in \binom{[n]}{k}} \tau_{S'\Delta S}\hat{f}(S')$ to obtain the eigenvalue of V_k'.

From Claim 9.1.5 and Claim 9.2.2, we separate the summation of $\tau_0 = \sum_{i=0}^{d-k} \binom{n-k}{i} \cdot \alpha_{k,k+i} \cdot \delta_i$ to $\sum_{even\ i} \binom{n-k}{i} \cdot \alpha_{k,k+i} \cdot \delta_i + \sum_{odd\ i} \binom{n-k}{i} \cdot \alpha_{k,k+i} \cdot \delta_i$. We replace δ_i and $\alpha_{k,k+i}$ by

the bound in Claim 9.1.5 and Claim 9.2.2:

$$\sum_{even\ i}\binom{n-k}{i}(-1)^{\frac{i}{2}+\frac{i}{2}}\left(\frac{(i-1)!!}{n^{\frac{i}{2}}}\frac{(i-1)!!}{n^{\frac{i}{2}}}+O(n^{-i-1})\right)$$
$$+\sum_{odd\ i}\binom{n-k}{i}(-1)^{\frac{i+1}{2}+\frac{i+1}{2}}\cdot O(n^{-\frac{i+1}{2}})\cdot O(n^{-\frac{i+1}{2}}).$$

It shows $\tau_0=\sum_{even\ i=0}^{d-k}\frac{(i-1)!!(i-1)!!}{i!}+O(n^{-1})$. For τ_{2l}, we bound it by $O(n^{-l})$ through similar method, where $O(n^{-l})$ comes from the fact that $\alpha_{k,k+i}=O(n^{-\frac{i}{2}})$, $\binom{n-k-2l}{i-t}<n^{i-t}$, and $\delta_{2l+i-2t}=O(n^{-\frac{2l+i-2t}{2}})$.

At last, we show the eigenvalue of V_k' is $O(1/n)$ close to τ_0, which is enough to finish the proof. From the fact that for any $T'\in\binom{[n]}{k-1}$, $\sum_{j\notin T'}\hat{f}(T'\cup j)=0$, we have (recall that$|S|=k$)

$$(k-i)\left(\sum_{S_0\in\binom{S}{i}}\sum_{S_1\in\binom{[n]\setminus S}{k-i}}\hat{f}(S_0\cup S_1)\right)+(i+1)\left(\sum_{S_0\in\binom{S}{i+1}}\sum_{S_1\in\binom{[n]\setminus S}{k-i-1}}\hat{f}(S_0\cup S_1)\right)$$
$$=\sum_{S_0\in\binom{S}{i}}\sum_{S_1\in\binom{[n]\setminus S}{k-i-1}}\sum_{j\notin S_0\cup S_1}\hat{f}(S_0\cup S_1\cup j)=0.$$

Thus we apply it on $\sum_i\tau_{2i}\big(\sum_{S'\in\binom{[n]}{k}:|S\cap S'|=k-i}\hat{f}(S')\big)$ to remove all $S'\neq S$. Let $\tau_{2k}'=\tau_{2k}$ and

$$\tau_{2k-2i-2}'=\tau_{2k-2i-2}-\frac{i+1}{k-i}\cdot\tau_{2k-2i}'.$$

Using the above rule, it is straightforward to verify

$$\sum_{i=j}^{k}\tau_{2i}\left(\sum_{S'\in\binom{[n]}{k}:|S\cap S'|=k-i}\hat{f}(S')\right)=\tau_{2j}'\left(\sum_{S'\in\binom{[n]}{k}:|S\cap S'|=k-j}\hat{f}(S')\right)$$

from $j=k$ to $j=0$ by induction. Therefore $\sum_{i=0}^{d}\tau_{2i}\big(\sum_{S'\in\binom{[n]}{k}:|S\cap S'|=k-i}\hat{f}(S')\big)=\tau_0'\hat{f}(S)$. Because $\tau_{2i}=O(n^{-i})$, we have $\tau_0'=\tau_0\pm O(1/n)$. (Remark: actually, $\tau_0'=\sum_{i=0}^{k}\tau_{2i}\cdot(-1)^i\binom{k}{i}$.)

From all discussion above, the eigenvalue of V_k' in A is τ_0', which is $\sum_{even\ i=0}^{d-k} \frac{(i-1)!!(i-1)!!}{i!} \pm O(n^{-1})$.

For V_k' in B of $k \geq 1$, observe that the difference $\delta_S \cdot \delta_T$ between $A(S,T)$ and $B(S,T)$ will not change the calculation of τ, because $\sum_{T \in \binom{[n]}{i}} \delta_S \delta_T \hat{f}(T) = \delta_S \delta_i \big(\sum_{T \in \binom{[n]}{i}} \hat{f}(T) \big) = 0$ from the fact f is orthogonal to V_0'. □

Because $\frac{(i-1)!!(i-1)!!}{i!} \leq 1$ for any even integer $i \geq 0$, we have the following two corollaries.

Corollary 9.2.4. *All non-zero eigenvalues of $\mathbb{E}_{D_p}[f^2]$ in the linear space of $\mathsf{span}\{\phi_S | S \in \binom{[n]}{\leq d}\}$ are between .5 and $[\frac{d}{2}] + 1 \leq d$.*

Corollary 9.2.5. *All non-zero eigenvalues of $\mathsf{Var}_{D_p}[f]$ in the linear space of $\mathsf{span}\{\phi_S | S \in \binom{[n]}{1,\cdots,d}\}$ are between .5 and $[\frac{d+1}{2}] \leq d$.*

Because $f + (\sum_i \phi_i)h \equiv f$ over $\mathrm{supp}(D_p)$ for any h of degree-at-most $d-1$, we define the projection of f onto V_{null}' to compare $\|f\|_2^2$ and $\mathbb{E}_{D_p}[f^2]$.

Definition 9.2.6. *Fix the global cardinality constraint $\sum_i x_i = (1-2p)n$ and the Fourier transform ϕ_S, let h_f denote the projection of a degree d multilinear polynomial f onto the null space $\mathsf{span}\{(\sum_i \phi_i)\phi_S | S \in \binom{[n]}{\leq d-1}\}$ of $E_{D_p}[f^2]$ and $\mathsf{Var}_{D_p}(f)$, i.e., $f - (\sum_i \phi_i)h_f$ is orthogonal to the eigenspace of an eigenvalue 0 in $\mathbb{E}_{D_p}[f^2]$ and $\mathsf{Var}_{D_p}(f)$.*

From the above two corollaries and the definition of h_f, we bound $\mathbb{E}_{D_p}[f^2]$ by $\|f\|_2^2$ as follows. For $\mathsf{Var}_{D_p}(f)$, we exclude $\hat{f}(\emptyset)$ because $\mathsf{Var}_{D_p}(f)$ is independent with $\hat{f}(\emptyset)$. Recall that $\|f\|_2^2 = E_{U_p}[f^2] = \sum_S \hat{f}(S)^2$.

Corollary 9.2.7. *For any degree d multilinear polynomial f and a global cardinality constraint $\sum_i x_i = (1-2p)n$, $\mathbb{E}_{D_p}[f^2] \leq d\|f\|_2^2$ and $\mathbb{E}_{D_p}[f^2] \geq 0.5\|f - (\sum_i \phi_i)h_f\|_2^2$.*

Corollary 9.2.8. *For any degree d multilinear polynomial f and a global cardinality constraint $\sum_i x_i = (1-2p)n$, $\mathsf{Var}_{D_p}(f) \leq d\|f - \hat{f}(\emptyset)\|_2^2$ and $\mathsf{Var}_{D_p}(f) \geq 0.5\|f - \hat{f}(\emptyset) - (\sum_i \phi_i)h_{f-\hat{f}(\emptyset)}\|_2^2$.*

9.3 Parameterized algorithm for CSPs above average with the bisection constraint

We prove that CSPs above average with the bisection constraint are fixed-parameter tractable. Given an instance $\mathcal{I}$ from d-ary CSPs and the bisection constraint $\sum_i x_i = 0$, we use the standard basis $\{\chi_S | S \in \binom{[n]}{\leq d}\}$ of the Fourier transform in U and abbreviate $f_{\mathcal{I}}$ to f. Recall that $\|f\|_2^2 = E_U[f^2] = \sum_S \hat{f}(S)^2$ and D is the uniform distribution on all assignments in $\{\pm 1\}^n$ complying with the bisection constraint.

For f with a small variance in D, we use $h_{f-\hat{f}(\emptyset)}$ to denote the projection of $f - \hat{f}(\emptyset)$ onto the null space $\mathsf{span}\{(\sum_i x_i)\chi_S | S \in \binom{[n]}{\leq d-1}\}$. We know $\|f - \hat{f}(\emptyset) - (\sum_i x_i)h_{f-\hat{f}(\emptyset)}\|_2^2 \leq 2\mathsf{Var}_D(f)$ from Corollary 9.2.8, i.e., the lower bound of the non-zero eigenvalues in $\mathsf{Var}_D(f)$. Then we show how to round $h_{f-\hat{f}(\emptyset)}$ in Section 9.3.1 to a degree $d-1$ polynomial h with integral coefficients such that $\|f - \hat{f}(\emptyset) - (\sum_i x_i)h\|_2^2 = O(\|f - \hat{f}(\emptyset) - (\sum_i x_i)h_{f-\hat{f}(\emptyset)}\|_2^2)$, which indicates that $f - \hat{f}(\emptyset) - (\sum_i x_i)h$ has a small kernel under the bisection constraint.

Otherwise, for f with a large variance in D, we show the hypercontractivity in D that $\mathbb{E}_D[(f - \mathbb{E}_D[f])^4] = O(\mathbb{E}_D[(f - \mathbb{E}_D[f])^2]^2)$ in Section 9.3.2. From the fourth moment method, we know there exists α in the support of D satisfying $f(\alpha) \geq \mathbb{E}_D[f] + \Omega(\sqrt{\mathsf{Var}_D[f]^2})$. At last, we prove the main theorem in Section 9.3.3.

Theorem 9.3.1. *Given an instance $\mathcal{I}$ of a CSP problem of arity d and a parameter t, there is an algorithm with running time $O(n^{3d})$ that either finds a kernel on at most $C_d t^2$ variables or certifies that $OPT \geq AVG + t$ under the bisection constraint for a constant $C_d = 24d^2 \cdot 7^d \cdot 9^d \cdot 2^{2d} \cdot \big(d!(d-1)! \cdots 2!\big)^2$.*

9.3.1 Rounding

In this section, we show that for any polynomial f of degree d with integral coefficients, there exists an efficient algorithm to round h_f into an integral-coefficient polynomial h while it keeps $\|f - (\sum_i x_i)h\|_2^2 = O(\|f - (\sum_i x_i)h_f\|_2^2)$.

Theorem 9.3.2. *For any constants γ and d, given a degree d multilinear polynomial f with $\|f - (\sum_i x_i)h_f\|_2^2 \leq \sqrt{n}$ whose Fourier coefficient $\hat{f}(S)$ is a multiple of γ for all $S \in \binom{[n]}{\leq d}$, there exists an efficient algorithm to find a degree-at-most $d-1$ polynomial h such that*

1. *The Fourier coefficients of h are multiples of $\frac{\gamma}{d!(d-1)!\cdots 2!}$, which demonstrates that the Fourier coefficients of $f - (\sum_i x_i)h$ are multiples of $\frac{\gamma}{d!(d-1)!\cdots 2!}$.*
2. *$\|f - (\sum_i x_i)h\|_2^2 \leq 7^d \cdot \|f - (\sum_i x_i)h_f\|_2^2$.*

The high level idea of the algorithm is to round $\hat{h}_f(S)$ to $\hat{h}(S)$ from the coefficients of weight $d-1$ to the coefficient of weight 0. At the same time, we guarantee that for any $k < d$, the rounding on the coefficients of weight k will keep $\|f - (\sum_i x_i)h\|_2^2 = O(\|f - (\sum_i x_i)h_f\|_2^2)$ in the same order.

Because h_f contains non-zero coefficients up to weight $d-1$, we first prove that we could round $\{\hat{h}_f(S) | S \in \binom{[n]}{d-1}\}$ to multiples of $\gamma/d!$. Observe that for $T \in \binom{[n]}{d}$, the coefficient of χ_T in $f - (\sum_i x_i)h_f$ is $\hat{f}(T) - \sum_{j \in T} \hat{h}_f(T \setminus j)$. Because $\sum_{T \in \binom{[n]}{d}} (\hat{f}(T) - \sum_{j \in T} \hat{h}_f(T \setminus j))^2 = o(n)$, $\hat{f}(T) - \sum_{j \in T} \hat{h}_f(T \setminus j)$ is close to 0 for most T in $\binom{[n]}{d}$. Hence $\sum_{j \in T} \hat{h}_f(T \setminus j) \mod \gamma$ is close to 0 for most T. Our start point is to prove that for any $S \in \binom{[n]}{d-1}$, $\hat{h}(S)$ is close to a multiple of $\gamma/d!$ from the above discussion.

Lemma 9.3.3. *If $\hat{f}(T)$ is a multiple of γ and $\hat{f}(T) - \sum_{S \in \binom{T}{d-1}} \hat{h}_f(S) = 0$ for all $T \in \binom{[n]}{d}$, then $\hat{h}_f(S)$ is a multiple of $\gamma/d!$ for all $S \in \binom{[n]}{d-1}$.*

Proof. From the two conditions, we know

$$\sum_{S\in\binom{T}{d-1}} \hat{h}_f(S) \equiv 0 \mod \gamma$$

for any $T \in \binom{[n]}{d}$. We prove that

$$(d-1)! \cdot \hat{h}_f(S_1) + (-1)^d (d-1)! \cdot \hat{h}_f(S_2) \equiv 0 \mod \gamma$$

for any $S_1 \in \binom{[n]}{d-1}$ and $S_2 \in \binom{[n]\setminus S}{d-1}$. Thus

$$0 \equiv (d-1)! \cdot \sum_{S_2\in\binom{T}{d-1}} \hat{h}_f(S_2) \equiv d! \cdot \hat{h}_f(S_1) \mod \gamma,$$

for any T with $S_1 \cap T = \emptyset$, which indicates $\hat{h}_f(S_1)$ is a multiple of $\gamma/d!$.

Without loss of generality, we assume $S_1 = \{1, 2, \cdots, d-1\}$ and $S_2 = \{k_1, k_2, \cdots, k_{d-1}\}$. For a subset $T \in \binom{S_1\cup S_2}{d}$, because $\sum_{S\in\binom{T}{d-1}} \hat{h}_f(S) = \sum_{j\in T} \hat{h}_f(T \setminus j)$, we use (T) to denote the equation

$$\sum_{j\in T} \hat{h}_f(T \setminus j) \equiv 0 \mod \gamma \tag{T}$$

Let $\beta_{d-1,1} = (d-2)!$ and $\beta_{d-i-1,i+1} = \frac{-i}{d-i-1} \cdot \beta_{d-i,i}$ for any $i \in \{1, \cdots, d-2\}$ (we choose $\beta_{d-1,1}$ to guarantee that all coefficients are integers). Consider the following linear combination of equations over $T \in \binom{S_1\cup S_2}{d}$ with coefficients $\beta_{d-i,i}$:

$$\sum_{i=1}^{d-1} \beta_{d-i,i} \sum_{T_1\in\binom{S_1}{d-i}, T_2\in\binom{S_2}{i}} (T_1 \cup T_2) \Rightarrow \sum_{i=1}^{d-1} \beta_{d-i,i} \sum_{T_1\in\binom{S_1}{d-i}, T_2\in\binom{S_2}{i}} \Big(\sum_{j\in T_1\cup T_2} \hat{h}_f(T_1 \cup T_2 \setminus j) \Big) \equiv 0 \mod \gamma. \tag{9.5}$$

Observe that for any $i \in \{1, \cdots, d-2\}$, $S \in \binom{S_1}{d-i-1}$, and $S' \in \binom{S_2}{i}$, the coefficient of $\hat{h}_f(S \cup S')$ is $i \cdot \beta_{d-i,i} + (d-i-1) \cdot \beta_{d-i-1,i+1} = 0$ in equation (9.5), where i comes from the number of choices of T_1 is $d-1-|S| = i$ and $d-i-1$ comes from the number of choices of T_2 is $(d-1)-|S'|$.

Hence equation (9.5) indicates that $(d-1)\beta_{d-1,1}\hat{h}_f(S_1) + (d-1)\beta_{1,d-1}\hat{h}_f(S_2) \equiv 0 \mod \gamma$. Setting into $\beta_{d-1,1} = (d-2)!$ and $\beta_{1,d-1} = (-1)^{d-2}(d-2)!$, we obtain

$$(d-1)! \cdot \hat{h}_f(S_1) + (-1)^d (d-1)! \cdot \hat{h}_f(S_2) \equiv 0 \mod \gamma.$$

□

Corollary 9.3.4. *If* $\sum_{T \in \binom{[n]}{d}} (\hat{f}(T) - \sum_{S \in \binom{T}{d-1}} \hat{h}_f(S))^2 = k = o(n^{0.6})$, *then for all* $S \in \binom{[n]}{d-1,}, \hat{h}_f(S)$ *is* $\frac{0.1}{d} \cdot \gamma/d!$ *close to a multiple of* $\gamma/d!$.

Proof. From the condition, we know that except for $n^{.8}$ choices of $T \in \binom{[n]}{d}$, $\sum_{S \in \binom{T}{d-1}} \hat{h}_f(S)$ is $n^{-.1}$ close to a multiple of γ because of $n^{.8} \cdot (n^{-.1})^2 > k$. Observe that the above proof depends on the Fourier coefficients in at most $2d+1$ variables of $S_1 \cup T$. Because $n^{0.8} = o(n)$, for any subset $S_1 \in \binom{[n]}{d-1}$, there is a subset $T \in \binom{[n] \setminus S_1}{d}$ such that for any $T' \in \binom{S_1 \cup T}{d}$, $\sum_{S \in \binom{T'}{d-1}} \hat{h}_f(S)$ is $n^{-.1}$ close to a multiple of γ.

Following the proof in Lemma 9.3.3, we obtain that $\hat{h}_f(S)$ is $\frac{(2d)!(d!)^2}{n^{.1}} < \frac{0.1}{d} \cdot \gamma/d!$ close to a multiple of $\gamma/d!$ for any $S \in \binom{[n]}{d-1}$. □

We consider a natural method to round h_f, which is to round $\hat{h}_f(S)$ to the closet multiple of $\gamma/d!$ for every $S \in \binom{[n]}{d-1}$.

Claim 9.3.5. *Let* h_{d-1} *be the rounding polynomial of* h_f *such that* $\hat{h}_{d-1}(S) = \hat{h}_f(S)$ *for any* $|S| \neq d-1$ *and* $\hat{h}_{d-1}(S)$ *is the closest multiple of* $\gamma/d!$ *to* $\hat{h}_f(S)$ *for any* $S \in \binom{[n]}{d-1}$. *Let* $\varepsilon(S) = \hat{h}_{d-1}(S) - \hat{h}_f(S)$.

If $|\varepsilon(S)| < .1/d \cdot \gamma/d!$ *and* $\alpha(T)$ *is a multiple of* γ *for any* T, *then*

$$\sum_{T \in \binom{[n]}{d}} \Big(\sum_{S \in \binom{T}{d-1}} \varepsilon(S) \Big)^2 \leq \sum_{T \in \binom{[n]}{d}} \Big(\alpha(T) - \sum_{S \in \binom{T}{d-1}} \hat{h}_f(S) \Big)^2.$$

Proof. For each $T \in \binom{[n]}{d}$, Because $\sum_{S\in\binom{T}{d-1}} |\varepsilon(S)| < 0.1 \cdot \gamma/d!$, then $|\sum_{S\in\binom{T}{d-1}} \varepsilon(S)| < |\alpha(T) - \sum_{S\in\binom{T}{d-1}} \hat{h}_f(S)|$. Hence we know $\sum_{T\in\binom{[n]}{d}} \left(\sum_{S\in\binom{T}{d-1}} \varepsilon(S)\right)^2 \leq \sum_{T\in\binom{[n]}{d}} \left(\alpha(T) - \sum_{S\in\binom{T}{d-1}} \hat{h}_f(S)\right)^2$. □

From now on, we use h_{d-1} to denote the degree $d-1$ polynomial of h_f after the above rounding process on the Fourier coefficients of weight $d-1$. Now we bound the summation of the square of the Fourier coefficients in $f - (\sum_i x_i)h_{d-1}$, i.e., $\|f - (\sum_i x_i)h_{d-1}\|_2^2$. Observe that rounding $\hat{h}_f(S)$ only affect the terms of $T \in \binom{[n]}{d}$ containing S and $T' \in \binom{[n]}{d-2}$ inside S, because $(\sum_i x_i)\hat{h}_f(S)\chi_S = \sum_{i\in S} \hat{h}_f(S)\chi_{S\setminus i} + \sum_{i\notin S} \hat{h}_f(S)\chi_{S\cup i}$.

Lemma 9.3.6. $\|f - (\sum_i x_i)h_{d-1}\|_2^2 \leq 7\|f - (\sum_i x_i)h_f\|_2^2$.

Proof. Let $\varepsilon(S) = \hat{h}_{d-1}(S) - \hat{h}_f(S)$. It is sufficient to prove

$$\sum_{T\in\binom{[n]}{d}} \left(\hat{f}(T) - \sum_{S\in\binom{T}{d-1}} \hat{h}_f(S) - \sum_{S\in\binom{T}{d-1}} \varepsilon(S)\right)^2 \leq 4 \sum_{T\in\binom{[n]}{d}} \left(\hat{f}(T) - \sum_{S\in\binom{T}{d-1}} \hat{h}_f(S)\right)^2, \quad (9.6)$$

and

$$\sum_{T'\in\binom{[n]}{d-2}} \left(\hat{f}(T') - \sum_{S\in\binom{T'}{d-3}} \hat{h}_f(S) - \sum_{j\notin T'} \hat{h}_f(T'\cup\{j\}) - \sum_{j\notin T'} \varepsilon(T'\cup\{j\})\right)^2 \leq 2\|f - (\sum_i x_i)h_f\|_2^2. \quad (9.7)$$

Equation (9.6) follows the fact that

$$\sum_{T\in\binom{[n]}{d}} \left(\sum_{S\in\binom{T}{d-1}} \varepsilon(S)\right)^2 \leq \sum_{T\in\binom{[n]}{d}} \left(\hat{f}(T) - \sum_{S\in\binom{T}{d-1}} \hat{h}_f(S)\right)^2$$

by Claim 9.3.5. From the inequality of arithmetic and geometric means, we know the cross terms:

$$\sum_{T\in\binom{[n]}{d}} 2 \cdot \left|\hat{f}(T) - \sum_{S\in\binom{T}{d-1}} \hat{h}_f(S)\right| \cdot \left|\sum_{S\in\binom{T}{d-1}} \varepsilon(S)\right| \leq 2 \sum_{T\in\binom{[n]}{d}} \left(\hat{f}(T) - \sum_{S\in\binom{T}{d-1}} \hat{h}_f(S)\right)^2.$$

For (9.7), observe that

$$\sum_{T'\in\binom{[n]}{d-2}}\Big(\sum_{j\notin T'}\varepsilon(T\cup\{j\})\Big)^2=(d-1)\sum_{S\in\binom{[n]}{d-1}}\varepsilon(S)^2+\sum_{S,S':|S\cap S'|=d-2}2\varepsilon(S)\varepsilon(S')$$
$$\leq\sum_{T\in\binom{[n]}{d}}\Big(\sum_{S\in\binom{T}{d-1}}\varepsilon(S)\Big)^2\leq\sum_{T\in\binom{[n]}{d}}\Big(\hat{f}(T)-\sum_{S\in\binom{T}{d-1}}\hat{h}_f(S)\Big)^2.$$

Hence we have

$$\sum_{T'\in\binom{[n]}{d-2}}\Big(\hat{f}(T')-\sum_{S\in\binom{T'}{d-3}}\hat{h}_f(S)-\sum_{j\notin T'}\hat{h}_f(T'\cup\{j\})\Big)+\sum_{T'\in\binom{[n]}{d-2}}\Big(\sum_{j\notin T'}\varepsilon(T\cup\{j\})\Big)^2$$
$$\leq\sum_{T'\in\binom{[n]}{d-2}}\Big(\hat{f}(T')-\sum_{S\in\binom{T'}{d-3}}\hat{h}_f(S)-\sum_{j\notin T'}\hat{h}_f(T'\cup\{j\})\Big)+\sum_{T\in\binom{[n]}{d}}\Big(\hat{f}(T)-\sum_{S\in\binom{T}{d-1}}\hat{h}_f(S)\Big)^2$$
$$\leq\|f-(\sum_i x_i)h_f\|_2^2.$$

We use the inequality of arithmetic and geometric means again to obtain inequality (9.7). □

Proof of Theorem 9.3.2. We apply Claim 9.3.5 and Lemma 9.3.6 for d times on the Fourier coefficients of h_f from $\{\hat{h}_f(S)|S\in\binom{[n]}{d-1}\},\{\hat{h}_f(S)|S\in\binom{[n]}{d-2}\},\cdots$ to $\{\hat{h}_f(S)|S\in\binom{[n]}{0}\}$ by choosing γ properly. More specific, let h_i be the polynomial after rounding the coefficients on $\binom{[n]}{\geq i}$ and $h_d=h_f$. Every time, we use Claim 9.3.5 to round coefficients of $\{\hat{h}_i(S)|S\in\binom{[n]}{i}\}$ from h_{i+1} for $i=d-1,\cdots,0$. We use different parameters of γ in different rounds: γ in the rounding of h_{d-1}, $\gamma/d!$ in the rounding of h_{d-2}, $\frac{\gamma}{d!\cdot(d-1)!}$ in the rounding of h_{d-3} and so on. After d rounds, all coefficients in h_0 are multiples of $\frac{\gamma}{d!(d-1)!(d-2)!\cdots 2!}$.

Because $\|f-(\sum_i x_i)h_i\|_2^2\leq 7\|f-(\sum_i x_i)h_{i+1}\|_2^2$ from Lemma 9.3.6. Eventually, $\|f-(\sum_i x_i)h_0\|_2^2\leq 7^d\cdot\|f-(\sum_i x_i)h_f\|_2^2$. □

9.3.2 $2\to 4$ Hypercontractive inequality under distribution D

We prove the $2\to 4$ hypercontractivity for a degree d polynomial g in this section.

Theorem 9.3.7. *For any degree-at-most d multilinear polynomial g, $\mathbb{E}_D[g^4] \leq 3d \cdot 9^{2d} \cdot \|g\|_2^4$.*

Recall that $\|g\|_2 = E_U[g^2]^{1/2} = (\sum_S \hat{g}(S)^2)^{1/2}$ and $g - (\sum_i x_i)h_g \equiv g$ in the support of D. Because $\|g - (\sum_i x_i)h_g\|_2^2 \leq 2\,\mathbb{E}_{x\sim D}[g^2]$ from the lower bound of non-zero eigenvalues in $E_D[g^2]$ in Corollary 9.2.4, without loss of generality, we assume g is orthogonal to the null space $\mathsf{span}\{(\sum_i x_i)\chi_S | S \in \binom{[n]}{\leq d-1}\}$.

Corollary 9.3.8. *For any degree-at-most d multilinear polynomial g, $\mathbb{E}_D[g^4] \leq 12d \cdot 9^{2d} \cdot \mathbb{E}_D[g^2]^2$.*

Before proving the above Theorem, we observe that uniform sampling a bisection $(S, \bar{S})$ is as same as first choosing a random perfect matching M and independently assigning each pair of M to the two subsets. For convenience, we use $P(M)$ to denote the product distribution on M and $\mathbb{E}_M$ to denote the expectation over a uniform random sampling of perfect matching M. Let $M(i)$ denote the vertex matched with i in M and $M(S) = \{M(i) | i \in S\}$. From the $2 \to 4$ hypercontractive inequality on product distribution $P(M)$, we have the following claim:

Claim 9.3.9. $\mathbb{E}_M[\mathbb{E}_{P(M)}[g^4]] \leq 9^d\,\mathbb{E}_M[\mathbb{E}_{P(M)}[g^2]^2]$.

Now we prove the main technical lemma of the $2 \to 4$ hypercontractivity under the bisection constraint to finish the proof.

Lemma 9.3.10. $\mathbb{E}_M[\mathbb{E}_{P(M)}[g^2]^2] \leq 3d \cdot 9^d \cdot \|g\|_2^4$.

Theorem 9.3.7 follows from Claim 9.3.9 and Lemma 9.3.10. Now we proceed to the proof of Lemma 9.3.10.

Proof of Lemma 9.3.10. Using $g(x) = \sum_{S \in \binom{[n]}{\leq d}} \hat{g}(S)\chi_S$, we rewrite $\mathbb{E}_M[\mathbb{E}_{P(M)}[g^2]^2]$ as

$$\mathop{\mathbb{E}}_{M}\left[\mathop{\mathbb{E}}_{P(M)}\left[\Big(\sum_{S\in\binom{[n]}{\leq d}}\hat{g}(S)\chi_S\Big)^2\right]^2\right]$$
$$= \mathop{\mathbb{E}}_{M}\left[\mathop{\mathbb{E}}_{P(M)}\Big[\sum_{S\in\binom{[n]}{\leq d}}\hat{g}(S)^2 + \sum_{S\in\binom{[n]}{\leq d},S'\in\binom{[n]}{\leq d},S'\neq S}\hat{g}(S)\hat{g}(S')\chi_{S\triangle S'}\Big]^2\right].$$

Notice that $\mathbb{E}_{P(M)}[\chi_{S\triangle S'}] = (-1)^{|S\triangle S'|/2}$ if and only if $M(S\triangle S') = S\triangle S'$; otherwise it is 0. We expand it to

$$\mathop{\mathbb{E}}_{M}\left[\left(\|g\|_2^2 + \sum_{S\in\binom{[n]}{\leq d},S'\in\binom{[n]}{\leq d},S'\neq S}\hat{g}(S)\hat{g}(S')\cdot 1_{S\triangle S'=M(S\triangle S')}\cdot(-1)^{|S\triangle S'|/2}\right)^2\right]$$
$$= \|g\|_2^4 + 2\|g\|_2^2\cdot\mathop{\mathbb{E}}_{M}\left[\sum_{S\in\binom{[n]}{\leq d},S'\in\binom{[n]}{\leq d},S'\neq S}\hat{g}(S)\hat{g}(S')\cdot 1_{S\triangle S'=M(S\triangle S')}\cdot(-1)^{|S\triangle S'|/2}\right]$$
$$+ \mathop{\mathbb{E}}_{M}\left[\left(\sum_{S\in\binom{[n]}{\leq d},S'\in\binom{[n]}{\leq d},S'\neq S}\hat{g}(S)\hat{g}(S')\cdot 1_{S\triangle S'=M(S\triangle S')}\cdot(-1)^{|S\triangle S'|/2}\right)^2\right].$$

We first bound the expectation of $\sum_{S\in\binom{[n]}{\leq d},S'\in\binom{[n]}{\leq d},S'\neq S}\hat{g}(S)\hat{g}(S')\cdot 1_{S\triangle S'=M(S\triangle S')}\cdot(-1)^{|S\triangle S'|/2}$ in the uniform distribution over all perfect matchings, then bound the expectation of its square. Observe that for a subset $U\subseteq[n]$ with even size, $\mathbb{E}_M[1_{U=M(U)}] = \frac{(|U|-1)(|U|-3)\cdots 1}{(m-1)(m-3)\cdots(m-|U|+1)}$ such that $\mathbb{E}_M[1_{U=M(U)}\cdot(-1)^{|U|/2}] = \delta_{|U|}$, i.e., the expectation $\mathbb{E}_D[\chi_U]$ of χ_U in D. Hence

$$\mathop{\mathbb{E}}_{M}\Big[\sum_{S\in\binom{[n]}{\leq d},S'\in\binom{[n]}{\leq d},S'\neq S}\hat{g}(S)\hat{g}(S')1_{S\triangle S'=M(S\triangle S')}\cdot(-1)^{|S\triangle S'|/2}\Big] \leq \sum_{S,S'}\hat{g}(S)\hat{g}(S')\cdot\delta_{S\triangle S'} = \mathop{\mathbb{E}}_{D}[g^2].$$

From Corollary 9.2.4, the largest non-zero eigenvalue of the matrix constituted by $\delta_{S\triangle S'}$ is at most d. Thus the expectation is upper bounded by $d\cdot\|g\|_2^2$.

We define g' to be a degree $2d$ polynomial $\sum_{T\in\binom{[n]}{\leq 2d}}\hat{g}'(T)\chi_T$ with coefficients

$$\hat{g}'(T) = \sum_{S\in\binom{[n]}{\leq d},S'\in\binom{[n]}{\leq d}:S\triangle S'=T}\hat{g}(S)\hat{g}(S')$$

for all $T \in \binom{[n]}{\leq 2d}$. Hence we rewrite

$$\begin{aligned}
&\mathop{\mathbb{E}}_{M}\left[\left(\sum_{S\in\binom{[n]}{\leq d},S'\in\binom{[n]}{\leq d},S'\neq S}\hat{g}(S)\hat{g}(S')\cdot 1_{S\triangle S'=M(S\triangle S')}\cdot(-1)^{|S\triangle S'|/2}\right)^2\right]\\
=&\mathop{\mathbb{E}}_{M}\left[\Big(\sum_{T\in\binom{[n]}{\leq 2d}}\hat{g}'(T)\cdot 1_{T=M(T)}(-1)^{|T|/2}\Big)^2\right]\\
=&\mathop{\mathbb{E}}_{M}\left[\sum_{T,T'}\hat{g}'(T)\hat{g}'(T')\cdot 1_{T=M(T)}1_{T'=M(T')}(-1)^{|T|/2+|T'|/2}\right].
\end{aligned}$$

Intuitively, because $|T| \leq 2d$ and $|T'| \leq 2d$, most of pairs T and T' are disjoint such that $\mathbb{E}_M[1_{T=M(T)}1_{T'=M(T')}] = \mathbb{E}_M[1_{T=M(T)}] \cdot \mathbb{E}_M[1_{T'=M(T')}]$. The summation is approximately $\mathbb{E}_M[\sum_T \hat{g}'(T)1_{T=M(T)}(-1)^{|T|/2}]^2$, which is bounded by $d^2\|g\|_2^4$ from the discussion above. However, we still need to bound the contribution from the correlated paris of T and T'.

Notice that $\|g'\|_2^2 = \mathbb{E}_U[g^4]$, which can be upper bounded by $\leq 9^d\|g\|_2^4$ from the standard $2 \to 4$ hypercontractivity.

Instead of bounding it by $\|g\|_2^4$ directly, we will bound it by $2d \cdot \|g'\|_2^2 \leq 2d \cdot 9^d\|g\|_2^4$ through the analysis on its eigenvalues and eigenspaces to this end. For convenience, we rewrite it to

$$\mathop{\mathbb{E}}_{M}\left[\sum_{T,T'}\hat{g}'(T)\hat{g}'(T)1_{T=M(T)}1_{T'=M(T')}(-1)^{|T|/2+|T'|/2}\right] = \sum_{T,T'}\hat{g}'(T)\hat{g}'(T')\Delta(T,T'),$$

where $\Delta(T,T') = 0$ if $|T|$ or $|T'|$ is odd, otherwise

$$\begin{aligned}
\Delta(T,T') = \mathop{\mathbb{E}}_{M}\left[1_{T\cap T'=M(T\cap T')}1_{T=M(T)}1_{T'=M(T')}(-1)^{|T|/2+|T'|/2}\right]&\\
= \frac{|T\cap T'-1|!!\cdot|T\setminus T'-1|!!\cdot|T'\setminus T-1|!!}{(n-1)(n-3)\cdots(n-|T\cup T'|+1)}(-1)^{|T\Delta T'|/2}.&
\end{aligned}$$

Let A' be the $\binom{n}{2d}\times\binom{n}{2d}$ matrix whose entry (T,T') is $\Delta(T,T')$. We prove that the eigenspace of A' with eigenvalue 0 is still $\mathsf{span}\{(\sum_i x_i)\chi_T | T \in \binom{[n]}{\leq 2d-1}\}$. Because $\Delta_{T,T'} \neq 0$ if and only

if $|T|, |T'|$,and $|T \cap T'|$ are even, it is sufficient to show $\sum_i A'(S, T\Delta i) = 0$ for all odd sized T and even sized S.

1. $|S \cap T|$ is odd: $\Delta(S, T \triangle i) \neq 0$ if and only if $i \in S$. We separate the calculation into $i \in S \cap T$ or not:

$$\sum_i A'(S, T \triangle i) = \sum_{i \in S \cap T} \Delta(S, T \setminus i) + \sum_{i \in S \setminus T} \Delta(S, T \cup i).$$

   Plugging in the definition of Δ, we obtain

$$\frac{|S \cap T| \cdot |S \cap T - 2|!! \cdot |S \setminus T|!! \cdot |T \setminus S|!!}{(n-1)(n-3)\cdots(n - |S \cup T| + 1)}(-1)^{|S|/2+|T-1|/2} + \frac{|S \setminus T| \cdot |S \cap T|!! \cdot |S \setminus T - 2|!! \cdot |T \setminus S|!!}{(n-1)(n-3)\cdots(n - |S \cup T| + 1)}(-1)^{|S|/2+|T+1|/2} = 0.$$

2. $|S \cap T|$ is even: $\Delta(S, T \triangle i) \neq 0$ if and only if $i \notin S$. We separate the calculation into $i \in T$ or not:

$$\sum_i A'(S, T \triangle i) = \sum_{i \in T \setminus S} \Delta(S, T \setminus i) + \sum_{i \notin S \cup T} \Delta(S, T \cup i).$$

   Plugging in the definition of Δ, we obtain

$$\frac{|T \setminus S| \cdot |S \cap T - 1|!! \cdot |S \setminus T - 1|!! \cdot |T \setminus S - 2|!!}{(n-1)(n-3)\cdots(n - |S \cup T| + 1)}(-1)^{|S|/2+|T-1|/2} + \frac{(n - |S \cup T|) \cdot |S \cap T - 1|!! \cdot |S \setminus T - 1|!! \cdot |T \setminus S|!!}{(n-1)(n-3)\cdots(n - |S \cup T|)}(-1)^{|S|/2+|T+1|/2} = 0.$$

From the same analysis in Section 9.2, the eigenspaces of A' are as same as the eigenspaces of A with degree $2d$ except the eigenvalues, whose differences are the differences between $\Delta_{S\triangle T}$ and $\delta_{S\triangle T}$. We can compute the eigenvalues of A' by the same calculation of eigenvalues in A. However, we bound the eigenvalues of A' by $0 \preceq A' \preceq A$ as follows.

Observe that for any S and T, $A'(S, T)$ and $A(S, T)$ always has the same sign. At the same time, $|A'(S, T)| = O(\frac{|A(S,T)|}{n^{|S \cap T|}})$. For a eigenspace V'_k in A, we focus on τ_0 because

the eigenvalue is $O(1/n)$-close to τ_0 from the proof of Theorem 9.2.3. We replace δ_i by any $\Delta(S,T)$ of $|S|=k, |T|=k+i$ and $|S\cap T|=i$ in $\tau_0=\sum_{i=0}^{d-k}\binom{n-k}{i}\cdot\alpha_{k,k+i}\cdot\delta_i$ to obtain τ_0' for A'. Thus $\alpha_{k,k+i}\cdot\Delta(S,T)=\Theta(\frac{\alpha_{k,k+i}\delta_i}{n^i})$ indicates $\tau_0'=\Theta(1)<\tau_0$ from the contribution of $i=0$. Repeat this calculation to τ_{2l}, we can show $\tau_{2l}'=O(\tau_{2l})$ for all l. Hence we know the eigenvalue of A' in V_k' is upper bounded by the eigenvalue of A from the cancellation rule of τ in the proof of Theorem 9.2.3. On the other hand, $A'\succeq 0$ from the definition that it is the expectation of a square term in M.

From Corollary 9.2.4 and all discussion above, we bound the largest eigenvalue of A' by $2d$. Therefore

$$\mathop{\mathbb{E}}_{M}\left[\sum_{T,T'}\hat{g}'(T)\hat{g}'(T)1_{T=M(T)}1_{T'=M(T')}(-1)^{|T|/2+|T'|/2}\right]$$
$$=\sum_{T,T'}\hat{g}'(T)\hat{g}'(T')\Delta(T,T')\le 2d\|g'\|_2^2\le 2d\cdot 9^d\cdot\|g\|_2^4.$$

Over all discussion above, $\mathbb{E}_M\left[\mathbb{E}_{P(M)}[g^2]^2\right]\le\|g\|_2^4+2d\|g\|_2^4+2d\cdot 9^d\cdot\|g\|_2^4\le 3d\cdot 9^d\cdot\|g\|_2^4$. □

9.3.3 Proof of Theorem 9.3.1

In this section, we prove Theorem 9.3.1. Let $f=f_{\mathcal{I}}$ be the degree d multilinear polynomial associated with the instance $\mathcal{I}$ and $g=f-\mathbb{E}_D[f]$ for convenience. We discuss $\mathsf{Var}_D[f]$ in two cases.

If $\mathsf{Var}_D[f]=\mathbb{E}_D[g^2]\ge 12d\cdot 9^{2d}\cdot t^2$, we have $\mathbb{E}_D[g^4]\le 12d\cdot 9^{2d}\cdot\mathbb{E}_D[g^2]^2$ from the $2\to 4$ hypercontractivity of Theorem 9.3.7. By Lemma 9.1.1, we know $\Pr_D[g\ge\frac{\sqrt{\mathbb{E}_D[g^2]}}{2\sqrt{12d\cdot 9^{2d}}}]>0$. Thus $\Pr_D[g\ge t]>0$, which demonstrates that $\Pr_D[f\ge\mathbb{E}_D[f]+t]>0$.

Otherwise we know $\mathsf{Var}_D[f]<12d\cdot 9^{2d}\cdot t^2$. We consider $f-\hat{f}(\emptyset)$ now. Let $h_{f-\hat{f}(\emptyset)}$ be the projection of $f-\hat{f}(\emptyset)$ onto the linear space such that $\|f-\hat{f}(\emptyset)-(\sum_i x_i)h_{f-\hat{f}(\emptyset)}\|_2^2\le 2\mathsf{Var}(f)$

from Corollary 9.2.8 and $\gamma = 2^{-d}$. From Theorem 9.3.2, we could round $h_{f-\hat{f}(\emptyset)}$ to h for $f - \hat{f}(\emptyset)$ in time $n^{O(d)}$ such that

1. the coefficients of $f - \hat{f}(\emptyset) - (\sum_i x_i)h$ are multiples of $\frac{\gamma}{d!(d-1)!\cdots 2!}$;
2. $\|f - \hat{f}(\emptyset) - (\sum_i x_i)h\|_2^2 \leq 7^d \|f - \hat{f}(\emptyset) - (\sum_i x_i)h_{f-\hat{f}(\emptyset)}\|_2^2 \leq 7^d \cdot 2 \cdot 12d \cdot 9^d \cdot t^2$.

We first observe that $f(\alpha) = f(\alpha) - (\sum_i \alpha_i)h(\alpha)$ for any α in the support of D. Then we argue $f - \hat{f}(\emptyset) - (\sum_i x_i)h$ has a small kernel, which indicates that f has a small kernel. From the above two properties, we know there are at most $\|f - \hat{f}(\emptyset) - (\sum_i x_i)h\|_2^2 / (\frac{\gamma}{d!(d-1)!\cdots 2!})^2$ non-zero coefficients in $f - \hat{f}(\emptyset) - (\sum_i x_i)h$. Because each of the nonzero coefficients contains at most d variables, the instance $\mathcal{I}$ has a kernel of at most $24d^2 \cdot 7^d \cdot 9^d \cdot 2^{2d} \cdot \big(d!(d-1)! \cdots 2!\big)^2 t^2$ variables.

The running time of this algorithm is the running time to find h_f and the rounding time $O(n^d)$. Therefore this algorithm runs in time $O(n^{3d})$.

9.4 $2 \to 4$ Hypercontractive inequality under distribution D_p

In this section, we prove the $2 \to 4$ hypercontractivity of low-degree multilinear polynomials in the distribution D_p conditioned on the global cardinality constraint $\sum_i x_i = (1-2p)n$.

We assume p is in $(0,1)$ such that $p \cdot n$ is a integer. Then we fix the Fourier transform to be the p-biased Fourier transform in this section, whose basis is $\{\phi_S | S \in \binom{[n]}{\leq d}\}$. Hence we use $\phi_1, \cdots, \phi_n$ instead of $x_1, \cdots, x_n$ and say that a function only depends on a subset of characters $\{\phi_i | i \in S\}$ if this function only takes input from variables $\{x_i | i \in S\}$. For a degree d multilinear polynomial $f = \sum_{S \in \binom{[n]}{\leq d}} \hat{f}(S)\phi_S$, we use $\|f\|_2 = E_{U_p}[f^2]^{1/2} = (\sum_S \hat{f}(S)^2)^{1/2}$ in this section.

We rewrite the global cardinality constraint as $\sum_i \phi_i = 0$. For convenience, we use $n_+ = (1-p)n$ to denote the number of $\sqrt{\frac{p}{1-p}}$'s in the global cardinality constraint of ϕ_i and $n_- = pn$ to denote the number of $-\sqrt{\frac{1-p}{p}}$'s. If $\frac{n_+}{n_-} = \frac{1-p}{p} = \frac{p_1}{p_2}$ for some integers p_1 and p_2, we could follow the approach in Section 9.3.2 that first partition $[n_+ + n_-]$ into tuples of size $p_1 + p_2$ then consider the production distribution over tuples. However, this approach will introduce a dependence on $p_1 + p_2$ to the bound, which may be superconstant. Instead of partitioning, we use induction on the number of characters and degree to prove the $2 \to 4$ hypercontractivity of low-degree multilinear polynomials in D_p.

Theorem 9.4.1. *For any degree-at-most d multilinear polynomial f on $\phi_1, \cdots, \phi_n$,*

$$\mathop{\mathbb{E}}_{D_p}[f(\phi_1, \ldots, \phi_n)^4] \leq 3 \cdot d^{3/2} \cdot \left(256 \cdot \left((\frac{1-p}{p})^2 + (\frac{p}{1-p})^2\right)^2\right)^d \cdot \|f\|_2^4.$$

Recall that h_f is the projection of f onto the null space $\mathsf{span}\{(\sum_i \phi_i)\phi_S | S \in \binom{[n]}{\leq d-1}\}$. We know $f - (\sum_i \phi_i)h_f \equiv f$ in $\mathrm{supp}(D_p)$, which indicates $\mathbb{E}_{D_p}[f^k] = \mathbb{E}_{D_p}[(f - (\sum_i \phi_i)h_f)^k]$ for any integer k. Without loss of generality, we assume f is orthogonal to $\mathsf{span}\{(\sum_i \phi_i)\phi_S | S \in \binom{[n]}{\leq d-1}\}$. From the lower bound of eigenvalues in $\mathbb{E}_{D_p}[f^2]$ by Corollary 9.2.4, $0.5\|f\|_2^2 \leq \mathbb{E}_{D_p}[f^2]$. We have a direct corollary as follows.

Corollary 9.4.2. *For any degree-at-most d multilinear polynomial f on $\phi_1, \cdots, \phi_n$,*

$$\mathop{\mathbb{E}}_{D_p}[f(\phi_1, \ldots, \phi_n)^4] \leq 12 \cdot d^{3/2} \cdot \left(256 \cdot \left((\frac{1-p}{p})^2 + (\frac{p}{1-p})^2\right)^2\right)^d \mathop{\mathbb{E}}_{D_p}[f(\phi_1, \ldots, \phi_n)^2]^2.$$

Note that since x_i can be written as a linear function of ϕ_i and linear transformation does not change the degree of the multilinear polynomial, we also have for any degree-at-most d multilinear polynomial g on $x_1, \cdots, x_n$,

$$\mathop{\mathbb{E}}_{D_p}[g(x_1, \ldots, x_n)^4] \leq 12 \cdot d^{3/2} \cdot \left(256 \cdot \left((\frac{1-p}{p})^2 + (\frac{p}{1-p})^2\right)^2\right)^d \mathop{\mathbb{E}}_{D_p}[g(x_1, \ldots, x_n)^2]^2.$$

Proof of Theorem 9.4.1. We assume the inequality holds for any degree $< d$ polynomials and use induction on the number of characters in a degree d multilinear polynomial f to prove that if the multilinear polynomial f of $\phi_1, \cdots, \phi_n$ depends on at most k characters of $\phi_1, \cdots, \phi_n$, then $\mathbb{E}_{D_p}[f^4] \leq d^{3/2} \cdot C^d \cdot \beta^k \cdot \|f\|_2^4$ for $C = 256 \cdot \left((\frac{1-p}{p})^2 + (\frac{p}{1-p})^2\right)^2$ and $\beta = 1 + 1/n$.

Base case. f is a constant function that is independent from $\phi_1, \cdots, \phi_n$, $\mathbb{E}_{D_p}[f^4] = \hat{f}(\emptyset)^4 = \|f\|_2^4$.

Induction step. Suppose there are $k \geq 1$ characters of $\phi_1, \cdots, \phi_n$ in f. Without loss of generality, we assume ϕ_1 is one of the characters in f and rewrite $f = \phi_1 h_0 + h_1$ for a degree $d-1$ polynomial h_0 with at most $k-1$ characters and a degree d polynomial h_1 with at most $k-1$ characters. Because f is a multilinear polynomial, $\|f\|_2^2 = \|h_0\|_2^2 + \|h_1\|_2^2$. We expand $\mathbb{E}_{D_p}[f^4] = \mathbb{E}_{D_p}[(\phi_1 h_0 + h_1)^4]$ to

$$\mathop{\mathbb{E}}_{D_p}[\phi_1^4 \cdot h_0^4] + 4 \mathop{\mathbb{E}}_{D_p}[\phi_1^3 \cdot h_0^3 \cdot h_1] + 6 \mathop{\mathbb{E}}_{D_p}[\phi_1^2 \cdot h_0^2 \cdot h_1^2] + 4 \mathop{\mathbb{E}}_{D_p}[\phi_1 \cdot h_0 \cdot h_1^3] + \mathop{\mathbb{E}}_{D_p}[h_1^4].$$

From the induction hypothesis, $\mathbb{E}_{D_p}[h_1^4] \leq C^d \beta^{k-1} \|h_1\|_2^4$ and

$$\mathop{\mathbb{E}}_{D_p}[\phi_1^4 \cdot h_0^4] \leq \max\{(\frac{1-p}{p})^2, (\frac{p}{1-p})^2\} \mathop{\mathbb{E}}_{D_p}[h_0^4] \leq d^{3/2} \cdot C^{d-0.5} \beta^{k-1} \|h_0\|_2^4. \tag{9.8}$$

Hence $\mathbb{E}_{D_p}[\phi_1^2 \cdot h_0^2 \cdot h_1^2] \leq (\mathbb{E}_{D_p}[\phi_1^4 h_0^4])^{1/2} (\mathbb{E}_{D_p}[h_1^4])^{1/2}$ from the Cauchy-Schwarz inequality. From the above discussion, this is at most

$$d^{3/4} \cdot C^{d/2} \cdot \beta^{(k-1)/2} \cdot \|h_1\|_2^2 \cdot d^{3/4} \cdot C^{(d-0.5)/2} \beta^{(k-1)/2} \|h_0\|_2^2 \leq d^{3/2} \cdot C^{d-1/4} \cdot \beta^{k-1} \cdot \|h_0\|_2^2 \|h_1\|_2^2. \tag{9.9}$$

Applying the inequality of arithmetic and geometric means on $\mathbb{E}_{D_p}[\phi_1^3 \cdot h_0^3 \cdot h_1]$, we know it is at most

$$(\mathop{\mathbb{E}}_{D_p}[\phi_1^4 h_0^4] + \mathop{\mathbb{E}}_{D_p}[\phi_1^2 h_0^2 h_1^2])/2 \leq d^{3/2} \big(C^{d-0.5} \beta^{k-1} \|h_0\|_2^4 + C^{d-1/4} \cdot \beta^{k-1} \cdot \|h_0\|_2^2 \|h_1\|_2^2\big)/2. \tag{9.10}$$

Finally, we bound $\mathbb{E}_{D_p}[\phi_1 \cdot h_0 \cdot h_1^3]$. However, we cannot apply the Cauchy-Schwarz inequality or the inequality of arithmetic and geometric means, because we cannot afford a term like $d^{3/2} \cdot C^d \beta^{k-1} \|h_1\|_2^4$ any more. We use $D_{\phi_1>0}$ ($D_{\phi_1<0}$ resp.) to denote the conditional distribution of D_p on fixing $\phi_1 = \sqrt{\frac{p}{1-p}}$ ($-\sqrt{\frac{1-p}{p}}$ resp.) and rewrite

$$\mathop{\mathbb{E}}_{D_p}[\phi_1 \cdot h_0 \cdot h_1^3] = \sqrt{p(1-p)} \mathop{\mathbb{E}}_{D_{\phi_1>0}}[h_0 \cdot h_1^3] - \sqrt{p(1-p)} \mathop{\mathbb{E}}_{D_{\phi_i<0}}[h_0 \cdot h_1^3]. \tag{9.11}$$

Let L be the matrix corresponding to the quadratic form $\mathbb{E}_{D_{\phi_1>0}}[fg] - \mathbb{E}_{D_{\phi_1<0}}[fg]$ for low-degree multilinear polynomials f and g (i.e. let L be a matrix such that $f^T L g = \mathbb{E}_{D_{\phi_1>0}}[fg] - \mathbb{E}_{D_{\phi_1<0}}[fg]$). The main technical lemma of this section is a upper bound on the spectral norm of L.

Lemma 9.4.3. *Let g be a degree $d (d \geq 1)$ multilinear polynomial on characters $\phi_2, \cdots, \phi_n$, we have*

$$|\mathop{\mathbb{E}}_{D_{\phi_1>0}}[g^2] - \mathop{\mathbb{E}}_{D_{\phi_1<0}}[g^2]| \leq \frac{3d^{3/2}}{p(1-p)} \cdot \frac{\|g\|_2^2}{\sqrt{n}}.$$

Therefore, the spectral norm of L is upper bounded by $\frac{3d^{3/2}}{p(1-p)} \cdot \frac{1}{\sqrt{n}}$.

From the above lemma, we rewrite the equation (9.11) from the upper bound of its eigenvalues:

$$\begin{aligned}\mathop{\mathbb{E}}_{D_p}[\phi_1 \cdot h_0 \cdot h_1^3] &= \sqrt{p(1-p)} \left(\mathop{\mathbb{E}}_{D_{\phi_1>0}}[h_0 \cdot h_1^3] - \mathop{\mathbb{E}}_{D_{\phi_1<0}}[h_0 \cdot h_1^3] \right) \\ &= \sqrt{p(1-p)} (h_0 h_1)^T L \cdot (h_1^2) \leq \sqrt{p(1-p)} \cdot \frac{3(2d)^{3/2}}{p(1-p)} \cdot \frac{1}{\sqrt{n}} \cdot \|h_0 h_1\|_2 \cdot \|h_1^2\|_2.\end{aligned}$$

Then we use the inequality of arithmetic and geometric means on it:

$$\mathop{\mathbb{E}}_{D_p}[\phi_1 \cdot h_0 \cdot h_1^3] \leq \sqrt{p(1-p)} \cdot \frac{10d^{3/2}}{p(1-p)} \cdot \frac{\|h_0 h_1\|_2^2 + \|h_1^2\|_2^2/n}{2}.$$

Next, we use the $2 \to 4$ hypercontractivity $\|h^2\|_2^2 = \mathbb{E}_{U_p}[h^4] \le 9^d \cdot \left(\frac{p^2}{1-p} + \frac{(1-p)^2}{p}\right)^d \|h\|_2^4$ in U_p and the Cauchy-Schwarz inequality to further simplify it to:

$$\begin{aligned}\mathop{\mathbb{E}}_{D_p}[\phi_1 \cdot h_0 \cdot h_1^3] &\le \frac{5d^{3/2}}{\sqrt{p(1-p)}} \cdot (\|h_0^2\|_2 \cdot \|h_1^2\|_2 + \frac{9^d \cdot \left(\frac{p^2}{1-p} + \frac{(1-p)^2}{p}\right)^d}{n}\|h_1\|_2^4) \\ &\le \frac{5d^{3/2}}{\sqrt{p(1-p)}} \cdot 9^d \cdot \left(\frac{p^2}{1-p} + \frac{(1-p)^2}{p}\right)^d (\|h_0\|_2^2 \cdot \|h_1\|_2^2 + \frac{1}{n}\|h_1\|_2^4). \quad (9.12)\end{aligned}$$

From all discussion, we bound $E[f^4]$ by the upper bound of each inequalities in (9.8),(9.10),(9.9),(9.12):

$$\begin{aligned}&E_{D_p}[(\phi_1 h_0 + h_1)^4] \\ =&E_{D_p}[\phi_1^4 \cdot h_0^4] + 4E_{D_p}[\phi_1^3 \cdot h_0^3 \cdot h_1] + 6E_{D_p}[\phi_1^2 \cdot h_0^2 \cdot h_1^2] + 4E_{D_p}[\phi_1 \cdot h_0 \cdot h_1^3] + E_{D_p}[h_1^4] \\ \le&d^{\frac{3}{2}} C^{d-\frac{1}{2}} \beta^{k-1} \|h_0\|_2^4 + 2d^{\frac{3}{2}} \left(C^{d-\frac{1}{2}} \beta^{k-1} \|h_0\|_2^4 + C^{d-\frac{1}{4}} \beta^{k-1} \|h_0\|_2^2 \|h_1\|_2^2\right) + 6d^{\frac{3}{2}} C^{d-\frac{1}{4}} \beta^{k-1} \|h_0\|_2^2 \|h_1\|_2^2 \\ &+ 4\frac{5d^{\frac{3}{2}}}{\sqrt{p(1-p)}} \cdot 9^d \cdot \left(\frac{p^2}{1-p} + \frac{(1-p)^2}{p}\right)^d \left(\|h_0\|_2^2 \|h_1\|_2^2 + \frac{1}{n}\|h_1\|_2^4\right) + d^{\frac{3}{2}} \cdot C^d \beta^{k-1} \|h_1\|_2^4 \\ \le&3d^{\frac{3}{2}} \cdot C^{d-\frac{1}{2}} \beta^{k-1} \|h_0\|_2^4 + \left(8d^{\frac{3}{2}} \cdot C^{d-\frac{1}{4}} \cdot \beta^{k-1} + \frac{20d^{\frac{3}{2}}}{\sqrt{p(1-p)}} \cdot 9^d \cdot \left(\frac{p^2}{1-p} + \frac{(1-p)^2}{p}\right)^d\right) \|h_0\|_2^2 \|h_1\|_2^2 \\ &+ \left(\frac{20d^{\frac{3}{2}}}{\sqrt{p(1-p)}} \cdot 9^d \cdot \left(\frac{p^2}{1-p} + \frac{(1-p)^2}{p}\right)^d \cdot \frac{1}{n} + d^{\frac{3}{2}} \cdot C^d \beta^{k-1}\right) \|h_1\|_2^4 \\ \le&d^{\frac{3}{2}} \cdot C^d \beta^k \|h_0\|_2^4 + d^{\frac{3}{2}} \cdot C^d \beta^k \cdot 2\|h_0\|_2^2 \|h_1\|_2^2 + d^{\frac{3}{2}} \cdot (C^d/n + C^d \beta^{k-1}) \|h_1\|_2^4 \le d^{\frac{3}{2}} \cdot C^d \beta^k \|f\|_2^4.\end{aligned}$$

□

We prove Lemma 9.4.3 to finish the proof. Intuitively, both $\mathbb{E}_{D_{\phi_1>0}}[g^2]$ and $\mathbb{E}_{D_{\phi_1<0}}[g^2]$ are close to $\mathbb{E}_{D_p}[g^2]$ (we add the dummy character ϕ_1 back in D_p) for a low-degree multilinear polynomial g; therefore their gap should be small compared to $\mathbb{E}_{D_p}[g^2] = O(\|g\|_2^2)$. Recall that D_p is a uniform distribution on the constraint $\sum_i \phi_i = 0$, i.e., there are always n_+

characters of ϕ_i with $\sqrt{\frac{p}{1-p}}$ and n_- characters with $-\sqrt{\frac{1-p}{p}}$. For convenience, we abuse the notation ϕ to denote a vector of characters $(\phi_1, \cdots, \phi_n)$.

Proof of Lemma 9.4.3. Let F be the p-biased distribution on $n-2$ characters with the global cardinality constraint $\sum_{i=2}^{n-1} \phi_i = -\sqrt{\frac{p}{1-p}} + \sqrt{\frac{1-p}{p}} = -q$, i.e., $n_+ - 1$ of the characters are always $\sqrt{\frac{p}{1-p}}$ and $n_- - 1$ of the characters are always $-\sqrt{\frac{1-p}{p}}$. Let $\vec{\phi}_{-i}$ denote the vector $(\phi_2, \cdots, \phi_{i-1}, \phi_{i+1}, \cdots, \phi_n)$ of $n-2$ characters such that we could sample $\vec{\phi}_{-i}$ from F. Hence $\phi \sim D_{\phi_1<0}$ is equivalent to the distribution that first samples i from $2, \cdots, n$ then fixes $\phi_i = \sqrt{\frac{p}{1-p}}$ and samples $\vec{\phi}_{-i} \sim F$. Similarly, $\phi \sim D_{\phi_1>0}$ is equivalent to the distribution that first samples i from $2, \cdots, n$ then fixes $\phi_i = -\sqrt{\frac{1-p}{p}}$ and samples $\vec{\phi}_{-i} \sim F$.

For a multilinear polynomial g depending on characters $\phi_2, \cdots, \phi_n$, we rewrite

$$\begin{aligned}
&\mathop{\mathbb{E}}_{D_{\phi_1>0}}[g^2] - \mathop{\mathbb{E}}_{D_{\phi_1<0}}[g^2] \\
&\qquad = \mathop{\mathbb{E}}_{i} \mathop{\mathbb{E}}_{\phi\sim F}\left[g(\phi_i = \sqrt{\frac{p}{1-p}}, \vec{\phi}_{-i} = \phi)^2 - g(\phi_i = -\sqrt{\frac{1-p}{p}}, \vec{\phi}_{-i} = \phi)^2\right]. \quad (9.13)
\end{aligned}$$

We will show its eigenvalue is upper bounded by $\frac{3d^{3/2}}{p(1-p)} \cdot \sqrt{1/n}$. We first use the Cauchy-Schwarz inequality:

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\phi\sim F}\left[\left(g(\phi_i = \sqrt{\frac{p}{1-p}}, \vec{\phi}_{-i} = \phi) - g(\phi_i = -\sqrt{\frac{1-p}{p}}, \vec{\phi}_{-i} = \phi)\right)\right. \\
&\qquad\qquad \left.\cdot\left(g(\phi_i = \sqrt{\frac{p}{1-p}}, \vec{\phi}_{-i} = \phi) + g(\phi_i = -\sqrt{\frac{1-p}{p}}, \vec{\phi}_{-i} = \phi)\right)\right] \\
&\le \mathop{\mathbb{E}}_{\phi\sim F}\left[\left(g(\phi_i = \sqrt{\frac{p}{1-p}}, \vec{\phi}_{-i} = \phi) - g(\phi_i = -\sqrt{\frac{1-p}{p}}, \vec{\phi}_{-i} = \phi)\right)^2\right]^{1/2} \\
&\qquad\qquad \cdot \mathop{\mathbb{E}}_{\phi\sim F}\left[\left(g(\phi_i = \sqrt{\frac{p}{1-p}}, \vec{\phi}_{-i} = \phi) + g(\phi_i = -\sqrt{\frac{1-p}{p}}, \vec{\phi}_{-i} = \phi)\right)^2\right]^{1/2}.
\end{aligned}$$

From the inequality of arithmetic and geometric means and the fact that $\mathbb{E}_{D_p}[g^2] \le d\|g\|_2^2$, observe that the second term is bounded by

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\phi\sim F}\left[\left(g(\phi_i = \sqrt{\frac{p}{1-p}}, \vec{\phi}_{-i} = \phi) + g(\phi_i = -\sqrt{\frac{1-p}{p}}, \vec{\phi}_{-i} = \phi)\right)^2\right] \\
\le& 2 \mathop{\mathbb{E}}_{D_{\phi_1>0}}\left[g(\phi_i = \sqrt{\frac{p}{1-p}}, \vec{\phi}_{-i} = \phi)^2\right] + 2 \mathop{\mathbb{E}}_{D_{\phi_1<0}}\left[g(\phi_i = -\sqrt{\frac{1-p}{p}}, \vec{\phi}_{-i} = \phi)^2\right] \\
\le& \frac{3}{p(1-p)} \mathop{\mathbb{E}}_{D_p}[g^2] \le \frac{3d}{p(1-p)} \cdot \|g\|_2^2.
\end{aligned}$$

Then we turn to $\mathbb{E}_{\phi\sim F}\left[\left(g(\phi_i = \sqrt{\frac{p}{1-p}}, \vec{\phi}_{-i} = \phi) - g(\phi_i = -\sqrt{\frac{1-p}{p}}, \vec{\phi}_{-i} = \phi)\right)^2\right]^{1/2}$. We use $g_i = \sum_{S:i\in S} \hat{f}(S)\phi_{S\setminus i}$ to replace $g(\phi_i = \sqrt{\frac{p}{1-p}}, \vec{\phi}_{-i} = \phi) - g(\phi_i = -\sqrt{\frac{1-p}{p}}, \vec{\phi}_{-i} = \phi)$:

$$\mathop{\mathbb{E}}_{\phi\sim F}\left[\left(\sum_{S:i\in S} \hat{g}(S)\left(\sqrt{\frac{p}{1-p}} + \sqrt{\frac{1-p}{p}}\right)\phi_{S\setminus i}\right)^2\right]^{1/2} = \left(\sqrt{\frac{p}{1-p}} + \sqrt{\frac{1-p}{p}}\right) \mathop{\mathbb{E}}_{\phi\sim F}\left[g_i(\phi)^2\right]^{1/2}.$$

Eventually we bound $\mathbb{E}_{\phi\sim F}[g_i(\phi)^2]$ by its eigenvalue. Observe that F is the distribution on $n-2$ characters with $(n_+ - 1)$ $\sqrt{\frac{p}{1-p}}$'s and $(n_- - 1)$ $-\sqrt{\frac{1-p}{p}}$'s, which indicates $\sum_j \phi_j + q = 0$ in F. However, the small difference between $\sum_j \phi_j + q = 0$ and $\sum_j \phi_j = 0$ will not change the major term in the eigenvalues of D_p. From the same analysis, the largest eigenvalue of $\mathbb{E}_F[g_i^2]$ is at most d. For completeness, we provide a calculation in Section 9.4.1.

Claim 9.4.4. *For any degree-at-most d multilinear polynomial g_i, $\mathbb{E}_{\phi\sim F}[g_i(\phi)^2] \le d\|g_i\|_2^2$.*

Therefore we have $\mathbb{E}_i[\mathbb{E}_{\phi\sim F}[g_i(\phi)^2]^{1/2}] \le \sqrt{d}\cdot\mathbb{E}_i[\|g_i(\phi)\|_2]$ and simplify the right hand

side of inequality (9.13) further:

$$\mathop{\mathbb{E}}_{i}\Bigg\{\mathop{\mathbb{E}}_{\phi\sim F}\left[\left(g(\phi_i=\sqrt{\frac{p}{1-p}},\vec{\phi}_{-i}=\phi)-g(\phi_i=-\sqrt{\frac{1-p}{p}},\vec{\phi}_{-i}=\phi)\right)^2\right]^{1/2}$$
$$\cdot\mathop{\mathbb{E}}_{\phi\sim F}\left[\left(g(\phi_i=\sqrt{\frac{p}{1-p}},\vec{\phi}_{-i}=\phi)+g(\phi_i=-\sqrt{\frac{1-p}{p}},\vec{\phi}_{-i}=\phi)\right)^2\right]^{1/2}\Bigg\}$$
$$\leq\mathop{\mathbb{E}}_{i}\left[\left(\sqrt{\frac{p}{1-p}}+\sqrt{\frac{1-p}{p}}\right)\cdot\sqrt{d}\cdot\|g_i\|_2\cdot\sqrt{\frac{3d}{p(1-p)}}\cdot\|g\|_2\right]$$
$$\leq\frac{3d}{p(1-p)}\cdot\|g\|_2\cdot\mathop{\mathbb{E}}_{i}[\|g_i\|_2]$$

Using the fact that $\sum_i\|g_i\|_2^2\leq d\|g\|_2^2$ and Cauchy-Schwartz again, we further simplify the expression above to obtain the desired upper bound on the absolute value of eigenvalues of $\mathbb{E}_{D_{\phi_1=1}}[g^2]-\mathbb{E}_{D_{\phi_1=-1}}[g^2]$:

$$\frac{3d}{p(1-p)}\cdot\|g\|_2\cdot\mathop{\mathbb{E}}_{i}[\|g_i\|_2]\leq\frac{3d}{p(1-p)}\cdot\|g\|_2\cdot\frac{(\sum_i\|g_i\|_2^2)^{1/2}}{\sqrt{n}}$$
$$\leq\frac{3d}{p(1-p)}\cdot\|g\|_2\cdot\sqrt{\frac{d}{n}}\|g\|_2\leq\frac{3d^{3/2}}{p(1-p)}\cdot\frac{\|g\|_2^2}{\sqrt{n}}.$$

□

9.4.1 Proof of Claim 9.4.4

We follow the approach in Section 9.2 to determine the eigenvalues of $\mathbb{E}_F[f^2]$ for a polynomial $f=\sum_{S\in\binom{[n-2]}{\leq d}}\hat{f}(S)\phi_S$. Recall that $\sum_i\phi_i+q=0$ over all support of F for $q=\frac{2p-1}{\sqrt{p(1-p)}}$ as defined before.

We abuse $\delta_k=\mathbb{E}_F[\phi_S]$ for a subset S with size k. We start with $\delta_0=1$ and $\delta_1=-q/(n-2)$. From $(\sum_i\phi_i+q)\phi_S\equiv 0$ for $k\leq d-1$ and any $S\in\binom{[n]}{k}$, we have

$$\mathbb{E}[\sum_{j\in S}\phi_j\phi_S+q\chi_S+\sum_{j\notin S}\phi_{S\cup j}]=0\Rightarrow k\cdot\delta_{k-1}+(k+1)q\cdot\delta_k+(n-2-k)\cdot\delta_{k+1}=0$$

Now we determine the eigenspaces of $\mathbb{E}_F[f^2]$. The eigenspace of an eigenvalue 0 is $\mathsf{span}\{(\sum_i \phi_i + q)\phi_S | S \in \binom{[n-2]}{\leq d-1}\}$. There are $d+1$ non-zero eigenspaces $V_0, V_1, \cdots, V_d$. The eigenspace V_i of $\mathbb{E}_F[f^2]$ is spanned by $\{\hat{f}(S)\phi_S | S \in \binom{[n-2]}{i}\}$. For each $f \in V_i$, f satisfies the following properties:

1. $\forall T' \in \binom{[n-2]}{i-1}, \sum_{j \notin T'} \hat{f}(S \cup j) = 0$ (neglect this contraint for V_0).
2. $\forall T \in \binom{[n]}{<i}, \hat{f}(T) = 0$.
3. $\forall T \in \binom{[n-2]}{>i}, \hat{f}(T) = \alpha_{k,|T|} \sum_{S \in T} \hat{f}(S)$ where $\alpha_{k,k} = 1$ and $\alpha_{k,k+i}$ satisfies

$$i \cdot \alpha_{k,k+i-1} + (k+i+1) \cdot q \cdot \alpha_{k,k+i} + (n-2-2k-i)\alpha_{k,k+i+1} = 0.$$

We show the calculation of $\alpha_{k,k+i}$ as follows: fix a subset T of size $k+i$ and consider the orthogonality between $(\sum_i \phi_i + q)]\phi_T$ and $f \in V_k$:

$$\begin{aligned}
&\sum_{j \in T} \alpha_{k,k+i-1} \sum_{S \in \binom{T \setminus j}{k}} \hat{f}(S) + (k+i+1) \cdot q \cdot \alpha_{k,k+i} \sum_{S \in \binom{T}{k}} \hat{f}(S) + \sum_{j \notin T} \alpha_{k,k+i+1} \sum_{S \in \binom{T \cup j}{k}} \hat{f}(S) = 0 \\
\Rightarrow &\sum_{S \in \binom{T}{k}} \Big((k+i+1) \cdot q \cdot \alpha_{k,k+i} + (n-2-k-i)\alpha_{k,k+i+1} + i \cdot \alpha_{k,k+i-1} \Big) \hat{f}(S) \\
&\qquad + \sum_{T' \in \binom{T}{k-1}} \alpha_{k,k+i+1} \sum_{j \notin T} \hat{f}(T' \cup j) = 0.
\end{aligned}$$

Using the first property $\forall T' \in \binom{[n-2]}{i-1}, \sum_{j \notin T'} \hat{f}(S \cup j) = 0$ to remove all $S' \notin T$, we have

$$\sum_{S \in \binom{T}{k}} \big(i \cdot \alpha_{k,k+i-1} + (k+i+1) \cdot q \cdot \alpha_{k,k+i} + (n-2-2k-i)\alpha_{k,k+i+1} \big) \hat{f}(S) = 0$$

We calculate the eigenvalues of V_k following the approach in Section 9.2. Fix S and S' with $i = |S \triangle S'|$, we still use τ_i to denote the coefficients of $\hat{f}(S')$ in the expansion of

$\sum_T(\delta_{S\triangle T} - \delta_S\delta_T)\hat{f}(T)$. Observe that τ_i is as same as the definition in Section 9.2 in terms of δ and α_k :

$$\tau_0 = \sum_{i=0}^{d-k} \binom{n-2-k}{i} \cdot \alpha_{k,k+i} \cdot \delta_i, \qquad \tau_{2l} = \sum_{i=0}^{d-k} \alpha_{k,k+i} \sum_{t=0}^{i} \binom{l}{t} \binom{n-2-k-2l}{i-t} \delta_{2l+i-2t}.$$

Observe that the small difference between $(\sum_i \phi_i + q) \equiv 0$ and $\sum_i \phi_i \equiv 0$ only changes a little in the recurrence formulas of δ and α. For δ_{2i} and $\alpha_{k,k+2i}$ of an integer i, the major term is still determined by δ_{2i-2} and $\alpha_{k,k+2i-2}$. For δ_{2i+1} and $\alpha_{k,k+2i+1}$, they are still in the same order (the constant before n^{-i-1} will not change the order). Using the same induction on δ and α, we have

1. $\delta_{2i} = (-1)^i \frac{(2i-1)!!}{n^i} + O(n^{-i-1})$;
2. $\delta_{2i+1} = O(n^{-i-1})$;
3. $\alpha_{k,k+2i} = (-1)^i \frac{(2i-1)!!}{n^i} + O(n^{-i-1})$;
4. $\alpha_{k,k+2i+1} = O(n^{-i-1})$.

Hence $\tau_0 = \sum_{even\ i=0}^{d-k} \frac{(i-1)!!(i-1)!!}{i!} + O(n^{-1})$ and $\tau_{2l} = O(n^{-l})$. Follow the same analysis in Section 9.2, we know the eigenvalue of V_k is $\tau_0 \pm O(\tau_2) = \sum_{even\ i=0}^{d-k} \frac{(i-1)!!(i-1)!!}{i!} + O(n^{-1})$.

From all discussion above, the eigenvalues of $\mathbb{E}_F[f^2]$ is at most $[\frac{d}{2}] + 1 \leq d$.

9.5 Parameterized algorithm for CSPs above average with global cardinality constraints

We show that CSPs above average with the global cardinality constraint $\sum_i x_i = (1-2p)n$ are fixed-parameter tractable for any $p \in [p_0, 1-p_0]$ with an integer pn. We still use D_p to denote the uniform distribution on all assignments in $\{\pm 1\}^n$ complying with $\sum_i x_i = (1-2p)n$.

Without loss of generality, we assume $p < 1/2$ and $(1-2p)n$ is an integer. We choose the standard basis $\{\chi_S\}$ in this section instead of $\{\phi_S\}$, because the Fourier coefficients in $\{\phi_S\}$ can be arbitrary small for some $p \in (0,1)$.

Theorem 9.5.1. *For any constant $p_0 \in (0,1)$ and d, given an instance of d-ary CSP, a parameter t, and a parameter $p \in [p_0, 1-p_0]$, there exists an algorithm with running time $n^{O(d)}$ that either finds a kernel on at most $C \cdot t^2$ variables or certifies that $OPT \geq AVG + t$ under the global cardinality constraint $\sum_i x_i = (1-2p)n$ for a constant $C = 160d^2 \cdot 30^d \left((\frac{1-p_0}{p_0})^2 + (\frac{p_0}{1-p_0})^2 \right)^d \cdot (d!)^{3d^2} \cdot (1/2p_0)^{4d}$.*

9.5.1 Rounding

Let f be a degree d polynomial whose coefficients are multiples of γ in the standard basis $\{\chi_S | S \in \binom{[n]}{\leq d}\}$. We show how to find an integral-coefficients polynomial h such that $f - \big(\sum_i x_i - (1-2p)n\big)h$ only depends on $O(\mathsf{Var}_{D_p}(f))$ variables. We use the rounding algorithm in Section 9.3.1 as a black box, which provides a polynomial h such that $f - \big(\sum_i x_i\big)h$ only depends on $O(\mathsf{Var}_D(f))$ variables (where D is the distribution conditioned on the bisection constraint). Without loss of generality, we assume $\hat{f}(\emptyset) = 0$ because $\mathsf{Var}_{D_p}(f)$ is independent with $\hat{f}(\emptyset)$.

Before proving that f depends on at most $O(\mathsf{Var}_{D_p}(f))$ variables, we first define the inactivity of a variable x_i in f.

Definition 9.5.2. *A variable x_i for $i \in [n]$ is inactive in $f = \sum_S \hat{f}(S)\chi_S$ if $\hat{f}(S) = 0$ for all S containing x_i.*
A variable x_i is inactive in f under the global cardinality constraint $\sum_i x_i = (1-2p)n$ if there exists a polynomial h such that x_i is inactive in $f - \big(\sum_i x_i - (1-2p)n\big)h$.

In general, there are multiple ways to choose h to turn a variable into inactive. However, if we know a subset S of d variables and the existence of some h to turn S

into inactive in $f - \big(\sum_i x_i - (1-2p)n\big)h$, we show that h is uniquely determined by S. Intuitively, for any subset S_1 with $d-1$ variables, there are $\binom{S_1 \cup S}{d} = \binom{2d-1}{d}$ ways to choose a subset of size d. Any d-subset T in $S_1 \cup S$ contains at least one inactive variable such that $\hat{f}(T) - \sum_{j \in T} \hat{h}(T \setminus j) = 0$ from the assumption. At the same time, there are at most $\binom{2d-1}{d-1}$ coefficients of $\hat{h}$ in $S_1 \cup S$. So there is only one solution of coefficients in $\hat{h}$ to satisfy these $\binom{2d-1}{d}$ equations.

Claim 9.5.3. *Given $f = \sum_{T \in \binom{[n]}{\leq d}} \hat{f}(T)\chi_T$ and $p \in [p_0, 1-p_0]$, let S be a subset with at least d variables such that there exists a degree $\leq d-1$ multilinear polynomial h turning S into inactive in $f - \big(\sum_i x_i - (1-2p)n\big)h$. Then h is uniquely determined by any d elements in S and f.*

Proof. Without lose of generality, we assume $S = \{1, \cdots, d\}$ and determine $\hat{h}(S_1)$ for $S_1 = \{i_1, \cdots, i_{d-1}\}$.

For simplicity, we first consider the case $S \cap S_1 = \emptyset$. From the definition, we know that for any $T \in \binom{S \cup S_1}{d}$, T contains at least one inactive variable, which indicates $\hat{f}(T) - \sum_{j \in T} \hat{h}(T \setminus j) = 0$. Hence we can repeat the argument in Lemma 9.3.3 to determine $\hat{h}(S_1)$ from $\hat{f}(T)$ over all $T \in \binom{S \cup S_1}{d}$.

Let $\beta_{d-1,1} = (d-2)!$ and $\beta_{d-i-1,i+1} = \frac{-i}{d-i-1}\beta_{d-i,i}$ for any $i \in \{1, \cdots, d-2\}$ be the parameters define in Lemma 9.3.3. For any $S_2 \in \binom{S}{d-1}$, by the same calculation,

$$\sum_{i=1}^{d-1} \beta_{d-i,i} \sum_{T_1 \in \binom{S_1}{d-i}, T_2 \in \binom{S_2}{i}} \left(\hat{f}(T_1 \cup T_2) - \sum_{j \in T_1 \cup T_2} \hat{h}(T_1 \cup T_2 \setminus j) \right) = 0$$

indicates that (all $\hat{h}(S)$ not S_1 or S_2 cancel with each other)

$$(d-1)! \cdot \hat{h}(S_1) + (-1)^d (d-1)! \cdot \hat{h}(S_2) = \sum_{i=1}^{d-1} \beta_{d-i,i} \sum_{T_1 \in \binom{S_1}{d-i}, T_2 \in \binom{S_2}{i}} \hat{f}(T_1 \cup T_2).$$

Hence $\hat{h}(S_2) = (-1)^{d-1}\hat{h}(S_1) + (-1)^d \sum_{i=1}^{d-1} \frac{\beta_{d-i,i}}{(d-1)!}\big(\sum_{T_1\in\binom{S_1}{d-i},T_2\in\binom{S_2}{i}} \hat{f}(T_1\cup T_2)\big)$ for any $S_2 \in \binom{S}{d-1}$. Replacing all $\hat{h}(S_2)$ in the equation $\hat{f}(S) - \sum_{S_2\in\binom{S}{d-1}} \hat{h}(S_2) = 0$, we obtain $\hat{h}(S_1)$ in terms of f and S.

If $S\cap S_1 \neq \emptyset$, then we set $S' = S\setminus S_1$. Next we add arbitrary $|S\cap S_1|$ more variables into S' such that $|S'| = d$ and $S'\cap S_1 = \emptyset$. Observe that $S'\cup S_1$ contains at least d inactive variables. Repeat the above argument, we could determine $\hat{h}(S_1)$.

After determining $\hat{h}(S_1)$ for all $S_1 \in \binom{[n]}{d-1}$, we repeat this argument for $S_1 \in \binom{[n]}{d-2}$ and so on. Therefore we could determine $\hat{h}(S_1)$ for all $S_1 \in \binom{[n]}{\leq d-1}$ from S and the coefficients in f. □

Remark 9.5.4. *The coefficients of h are multiples of $\gamma/d!$ if the coefficients of f are multiples of γ.*

Let h_1 and h_2 be two polynomials such that at least d variables are inactive in both $f-\big(\sum_i x_i-(1-2p)n\big)h_1$ and $f-\big(\sum_i x_i-(1-2p)n\big)h_2$. We know that $h_1 = h_2$ from the above claim. Furthermore, it implies that any variable that is inactive in $f-\big(\sum_i x_i-(1-2p)n\big)h_1$ is inactive in $f-\big(\sum_i x_i-(1-2p)n\big)h_2$ from the definition, and vice versa.

Based on this observation, we show how to find a degree $d-1$ function h such that there are fews active variables left in $f-\big(\sum_i x_i-(1-2p)n\big)h$. The high level is to random sample a subset Q of $(1-2p)n$ variables and restrict all variables in Q to 1. Thus the rest variables constitutes the bisection constraint on $2pn$ variables such that we could use the rounding process in Section 9.3.1. Let k be a large number, $Q_1,\cdots,Q_k$ be k random subsets and $h_1,\cdots,h_k$ be the k functions after rounding in Section 9.3.1. Intuitively, the number of active variables in $f-(\sum_{i\notin Q_1} x_i)h_1,\cdots,f-(\sum_{i\notin Q_i} x_k)h_k$ are small with high probability such that $h_1,\cdots,h_k$ share at least d inactive variables. We can use one function h to represent $h_1,\cdots,h_k$ from the above claim such that the union of inactive variables in

$f-(\sum_{i\notin Q_j}x_i)h_j$ over all $j\in[k]$ are inactive in $f-(\sum_i x_i)h$ from the definition. Therefore there are a few active variables in $f-(\sum_i x_i)h$.

Let us move to $f-\big(\sum_i x_i-(1-2p)n\big)h$. Because h is a degree-at-most $d-1$ function, $(1-2p)n\cdot h$ is a degree $\le d-1$ function. Thus we know that the number of active variables among degree d terms in $f-\big(\sum_i x_i-(1-2p)n\big)h$ is upper bounded by the number of active variables in $f-(\sum_i x_i)h$. For the degree $<d$ terms left in $f-\big(\sum_i x_i-(1-2p)n\big)h$, we repeat the above process again.

Theorem 9.5.5. *Given a global cardinality constraint $\sum_i x_i=(1-2p)n$ and a degree d function $f=\sum_{S\in\binom{[n]}{\le d}}\hat{f}(S)$ with $\mathsf{Var}_{D_p}(f)<n^{0.5}$ and coefficients of multiples of γ , there is an efficient algorithm running in time $O(dn^{2d})$ to find a polynomial h such that there are at most $\frac{C'_{p,d}\cdot\mathsf{Var}_{D_p}(f)}{\gamma^2}$ active variables in $f-\big(\sum_i x_i-(1-2p)n\big)h$ for $C'_{p,d}=\frac{20d^2 7^d\cdot(d!)^{2d^2}}{(2p)^{4d}}$.*

Proof. For any subset $Q\in\binom{[n]}{(1-2p)n}$, we consider the assignments conditioned on $x_Q=\vec{1}$ and use f_Q to denote the restricted function f on $x_Q=\vec{1}$. Conditioned on $x_Q=\vec{1}$, the global cardinality constraint on the rest variables is $\sum_{i\notin Q}x_i=0$. We use D_Q denote the distribution on assignments of $\{x_i|i\notin Q\}$ satisfying $\sum_{i\notin Q}x_i=0$, i.e., the distribution of $\{x_i|i\notin\bar{Q}\}$ under the bisection constraint.

Let $X_Q(i)\in\{0,1\}$ denote whether x_i is active in f_Q under the bisection constraint of $\bar{Q}$ or not after the bisection rounding in Theorem 9.3.2. From Theorem 9.3.2, we get an upper bound on the number of active variables in f_Q, i.e.,

$$\sum_i X_Q(i)\le 2C'_d\cdot\mathsf{Var}_{D_Q}(f_Q)$$

for $C'_d=\frac{7^d(d!(d-1)!\cdots 2!)^2}{\gamma^2}$ and any Q with $\mathsf{Var}_{D_Q}(f_Q)=O(n^{0.6})$.

We claim that

$$\mathop{\mathbb{E}}_{Q}[\mathsf{Var}_{D_Q}(f_Q)]\le\mathsf{Var}_{D_p}(f).$$

From the definition, $\mathbb{E}_Q[\mathsf{Var}_{D_Q}(f_Q)] = \mathbb{E}_Q\,\mathbb{E}_{y\sim D_Q}[f_Q(y)^2] - \mathbb{E}_Q\left[\mathbb{E}_{y\sim D_Q}[f_Q(y)]\right]^2$. At the same time, we observe that $\mathbb{E}_Q\,\mathbb{E}_{y\sim D_Q}[f_Q(y)^2] = \mathbb{E}_D[f^2]$ and $\mathbb{E}_Q[\mathbb{E}_{y\sim D_Q} f_Q(y)]^2 \geq \mathbb{E}_{D_p}[f]^2$. Therefore $\mathbb{E}_Q[\mathsf{Var}_{D_Q}(f_Q)] \leq \mathsf{Var}_{D_p}[f]$. One observation is that $\Pr_Q[\mathsf{Var}_{D_Q} \geq n^{0.6}] < n^{-0.1}$ from the assumption $\mathsf{Var}_{D_p}(f) < n^{0.5}$, which is very small such that we can neglect it in the rest of proof. From the discussion above, we have $\mathbb{E}_Q[\sum_i X_Q(i)] \leq 2C'_d \cdot \mathsf{Var}_{D_p}(f)$ with high probability.

Now we consider the number of i's with $\mathbb{E}_Q[X_Q(i)] \leq \frac{(2p)^{2d}}{5d}$. Without loss of generality, we use m to denote the number of i's with $\mathbb{E}_Q[X_Q(i)] \leq \frac{(2p)^{2d}}{5d}$ and further assume these variables are $\{1, 2, \cdots, m\}$ for convenience. Hence for any $i > m$, $\mathbb{E}_Q[X_Q(i)] > \frac{(2p)^{2d}}{5d}$. We know the probability $\mathsf{Var}_{D_Q}(f) \leq \frac{2\mathsf{Var}_{D_p}(f)}{(2p)^{2d}}$ is at least $1 - \frac{(2p)^{2d}}{2}$, which implies

$$n - m \leq \frac{2C'_d \cdot \mathsf{Var}_{D_Q}(f)}{\frac{(2p)^{2d}}{5d}} \leq \frac{20d \cdot C'_d \mathsf{Var}_{D_p(f)}}{(2p)^{4d}}.$$

We are going to show that $\mathbb{E}_Q[X_Q(i)]$ is either 0 or at least $\frac{(2p)^{2d}}{5d}$, which means that only $x_{m+1}, \cdots, x_n$ are active in f under $\sum_i x_i = 0$. Then we discuss how to find out a polynomial h_d such that $x_1, \cdots, x_m$ are inactive in the degree d terms of $f - \left(\sum_i x_i - (1-2p)n\right)h_d$.

We fix d variables $x_1, \cdots, x_d$ and pick d arbitrary variables $x_{j_1}, \cdots, x_{j_d}$ from $\{d+1, \cdots, n\}$. We focus on $\{x_1, x_2, \cdots, x_d, x_{j_1}, \cdots, x_{j_d}\}$ now. With probability at least $(2p)^{2d} - o(1) \geq 0.99(2p)^{2d}$ over random sampling Q, none of these $2d$ variables is in Q. At the same time, with probability at least $1 - 2d \cdot \frac{(2p)^{2d}}{5d}$, all variables in the intersction $\{x_1, \ldots, x_m\} \cap \{x_1, x_2, \cdots, x_d, x_{j_1}, \cdots, x_{j_d}\}$ are inactive in f_Q under the bisection constraint on $\bar{Q}$ ($2d$ is for $x_{j_1}, \cdots, x_{j_d}$ if necessary). Therefore, with probability at least $0.99(2p)^{2d} - 2d \cdot \frac{(2p)^{2d}}{5d} - \frac{(2p)^{2d}}{2} \geq 0.09(2p)^{2d}$, $x_1, x_2, \cdots, x_d$ are inactive in f_Q under $\sum_{i\notin Q} x_i = 0$ and $n - m$ is small. Namely there exists a polynomial $h_{x_{j_1},\cdots,x_{j_d}}$ such that the variables in $\{x_1, \ldots, x_m\} \cap \{x_1, x_2, \cdots, x_d, x_{j_1}, \cdots, x_{j_d}\}$ are inactive in $f_Q - (\sum_{i\notin Q} x_i)h_{x_{j_1},\cdots,x_{j_d}}$.

Now we apply Claim 9.5.3 on $S = \{1, \cdots, d\}$ in f to obtain the unique polynomial h_d, which is the combination of $h_{x_{j_1},\cdots,x_{j_d}}$ over all choices of $j_1, \cdots, j_d$, and consider $f - (\sum_i x_i)h_d$.

Because of the arbitrary choices of $x_{j_1}, \cdots, x_{j_d}$, it implies that $x_1, \cdots, x_d$ are inactive in $f - (\sum_i x_i)h_d$. For example, we fix any $j_1, \cdots, j_d$ and $T = \{1, j_1, \cdots, j_{d-1}\}$. we know $\hat{f}(T) - \sum_{j \in T} \hat{h}_{x_{j_1}, \cdots, x_{j_d}}(T \setminus j) = 0$ from the definition of $h_{x_{j_1}, \cdots, x_{j_d}}$. Because h_d agrees with $h_{x_{j_1}, \cdots, x_{j_d}}$ on the Fourier coefficients from Claim 9.5.3, we have $\hat{f}(T) - \sum_{j \in T} \hat{h}_d(T \setminus j) = 0$.

Furthermore, it implies that $x_{d+1}, \cdots, x_m$ are also inactive in $f - (\sum_i x_i)h_d$. For example, we fix $j_1 \in \{d+1, \cdots, m\}$ and choose $j_2, \cdots, j_d$ arbitrarily. Then $x_1, \cdots, x_d$, and x_{j_1} are inactive in $f_Q - (\sum_{i \notin Q} x_i)h_{x_{j_1}, \cdots, x_{j_d}}$ for some Q from the discussion above, which indicates that x_{j_1} are inactive in $f - (\sum_i x_i)h_d$ by Claim 9.5.3.

To find h_d in time $O(n^{2d})$, we enumerate all possible choices of d variables in $[n]$ as S. Then we apply Claim 9.5.3 to find the polynomial h_S corresponding to S and check $f - (\sum_i x_i)h_S$. If there are more than m inactive variables in $f - (\sum_i x_i)h_S$, then we set $h_d = h_S$. Therefore the running time of this process is $\binom{n}{d} \cdot O(n^d) = O(n^{2d})$.

Hence, we can find a polynomial h_d efficiently such that at least m variables are inactive in $f - (\sum_i x_i)h_d$. Let us return to the original global cardinality constraint $\sum_i x_i = (1-2p)n$. Let

$$f_d = f - \big(\sum_i x_i - (1-2p)n\big)h_d.$$

$x_1, \cdots, x_m$ are no longer inactive in f_d because of the extra term $(1-2p)n \cdot h$. However, $x_1, \cdots, x_m$ are at least independent with the degree d terms in f_d. Let A_d denote the set for active variables in the degree d terms of f_d, which is less than $\frac{20d \cdot C'_d \mathsf{Var}_{D_p(f)}}{(2p)^{4d}}$ from the upper bound of $n - m$.

For f_d, observe that $\mathsf{Var}_{D_p}(f_d) = \mathsf{Var}_{D_p}(f)$ and all coefficients of f_d are multiples of $\gamma_{d-1} = \gamma/d!$ from Claim 9.5.3. For f_d, we neglect its degree d terms in A_d and treat it as a degree $d-1$ function from now on. Then we could repeat the above process again for the degree $d-1$ terms in f_d to obtain a degree $d-2$ polynomial h_{d-1} such that the active set A_{d-1} in the degree $d-1$ terms of $f_{d-1} = f_d - \big(\sum_i x_i - (1-2p)n\big)h_{d-1}$ contains at most

$\frac{20d\cdot C'_{d-1}\mathsf{Var}_{D_p(f)}}{(2p)^{4d}}$ variables for $C'_{d-1} = \frac{((d-1)^2\cdots 2!)^2}{\gamma_{d-1}^2}$. At the same time, observe that the degree of $\big(\sum_i x_i - (1-2p)n\big)h_{d-1}$ is at most $d-1$ such that it will not introduce degree d terms to f_{d-1}. Then we repeat it again for terms of degree $d-2, d-3$, and so on.

To summarize, we can find a polynomial h such that $A_d \cup A_{d-1} \cdots \cup A_1$ is the active set in $f - \big(\sum_i x_i - (1-2p)n\big)h$. At the same time, $|A_d \cup A_{d-1} \cdots \cup A_1| \leq \sum_i |A_i| \leq \frac{20d^2 7^d \cdot \mathsf{Var}_{D_p}(f)\cdot (d!)^{2d^2}}{\gamma^2\cdot(2p)^{4d}}$. □

9.5.2 Proof of Theorem 9.5.1

In this section, we prove Theorem 9.5.1. Let $f = f_{\mathfrak{I}}$ be the degree d function associated with the instance $\mathfrak{I}$ and $g = f - \mathbb{E}_{D_p}[f]$ for convenience. We discuss $\mathsf{Var}_{D_p}[f]$ in two cases.

If $\mathsf{Var}_{D_p}[f] = \mathbb{E}_{D_p}[g^2] \geq 8\left(16\cdot\left((\frac{1-p}{p})^2 + (\frac{p}{1-p})^2\right)\cdot d^3\right)^d \cdot t^2$, we have

$$\mathop{\mathbb{E}}_{D_p}[g^4] \leq 12\cdot\left(256\cdot\left((\frac{1-p}{p})^2 + (\frac{p}{1-p})^2\right)^2\cdot d^6\right)^d \mathop{\mathbb{E}}_{D_p}[g^2]^2$$

from the $2\to 4$ hypercontractivity in Theorem 9.4.2. By Lemma 9.1.1, we know

$$\Pr_{D_p}\left[g \geq \frac{\sqrt{\mathbb{E}_{D_p}[g^2]}}{2\sqrt{12\cdot\left(256\cdot\left((\frac{1-p}{p})^2 + (\frac{p}{1-p})^2\right)^2\cdot d^6\right)^d}}\right] > 0.$$

Thus $\Pr_{D_p}[g \geq t] > 0$, which demonstrates that $\Pr_{D_p}[f \geq \mathbb{E}_{D_p}[g] + t] > 0$.

Otherwise we know $\mathsf{Var}_{D_p}[f] \leq 8\left(16\cdot\left((\frac{1-p}{p})^2 + (\frac{p}{1-p})^2\right)\cdot d^3\right)^d \cdot t^2$. We set $\gamma = 2^{-d}$. From Theorem 9.5.5, we could find a degree $d-1$ function h in time $O(n^{2d})$ such that $f - \big(\sum_i x_i - (1-2p)n\big)h$ contains at most $\frac{C'_{p,d}\cdot\mathsf{Var}_{D_p}(f)}{\gamma^2}$ variables. We further observe that $f(\alpha) = f(\alpha) - \big(\sum_i \alpha_i - (1-2p)n\big)h(\alpha)$ for any α in the support of D_p. Then we know the

kernel of f and $\mathfrak{I}$ is at most

$$8\left(16\cdot\left((\frac{1-p}{p})^2+(\frac{p}{1-p})^2\right)\cdot d^3\right)^d\cdot t^2\cdot\frac{C'_{p,d}}{\gamma^2}$$
$$<8\left(16\cdot\left((\frac{1-p}{p})^2+(\frac{p}{1-p})^2\right)\cdot d^3\right)^d\cdot t^2\cdot\frac{20d^2 7^d\cdot \mathsf{Var}_{D_p}(f)\cdot(d!)^{2d^2}\cdot 2^{2d}}{\gamma^2\cdot(2p)^{4d}}<C\cdot t^2.$$

The running time of this algorithm is $O(dn^{2d})$.

Chapter 10

Integrality Gaps for Dispersers and Bipartite Expanders

In this chapter, we study the vertex expansion of bipartite graphs. For convenience, we always use G to denote a bipartite graph and $[N] \cup [M]$ to denote the vertex set of G. Let D and d denote the maximal degree of vertices in $[N]$ and $[M]$, respectively. For a subset S in $[N] \cup [M]$ of a bipartite graph $G = ([N], [M], E)$, we use $\Gamma(S)$ to denote its neighbor set $\{j | \exists i \in S, (i, j) \in E\}$. We consider the following two useful concepts in bipartite graphs:

Definition 10.0.1. *A bipartite graph $G = ([N], [M], E)$ is a (k, s)-disperser if for any subset $S \subseteq [N]$ of size k, the neighbor set $\Gamma(S)$ contains at least s distinct vertices.*

Definition 10.0.2. *A bipartite graph $G = ([N], [M], E)$ is a (k, a)-expander if for any subset $S \subseteq [N]$ of size k, the neighbor set $\Gamma(S)$ contains at least $a \cdot k$ distinct vertices. It is a $(\leq K, a)$ expander if it is a (k, a)-expander for all $k \leq K$.*

Because dispersers focus on hitting most vertices in $[M]$, and expanders emphasize that the expansion is in proportion of the degree D , it is often more convenient to use parameters ρ, δ, and ϵ for $k = \rho N, s = (1 - \delta)M$, and $a = (1 - \epsilon)D$ for dispersers and expanders.

These two combinatorial objects have wide applications in computer science. Dispersers are well known for obtaining non-trivial derandomization results, e.g., for derandomization of inapproximability results for MAX Clique and other NP-Complete problems

[Zuc96a, TZ04, Zuc07], deterministic amplification [Sip88], and oblivious sampling [Zuc96b]. Dispersers are also closely related to other combinatorial constructions such as randomness extractors, and some constructions of dispersers follow the constructions of randomness extractors directly [TZ04, BKS+10, Zuc07]. Explicit constructions achieving almost optimal degree have been designed by Ta-Shma [Ta-02] and Zuckerman [Zuc07], respectively, in different important parameter regimes.

For bipartite expanders, it is well known that the probabilistic method provides very good expanders, and some applications depend on the existence of such bipartite expanders, e.g., proofs of lower bounds in different computation models [Gri01a, BOT02]. Expanders also constitute an important part in other pseudorandom constructions, such as expander codes [SS96] and randomness extractors [TZ04, CRVW02, GUV09b]. A beautiful application of bipartite expanders was given by Buhrman et.al. [BMRV00] in the static membership problem (see [CRVW02] for more applications and the reference therein). Explicit constructions for expansion $a = (1-\epsilon)D$ with almost-optimal parameters have been designed in [CRVW02] and [TZ04, GUV09b] for constant degree and super constant degree respectively.

We consider the natural problem of how to approximate the vertex expansion of ρN-subsets in a bipartite graph G on $[N] \cup [M]$ in terms of the degrees D, d, and the parameter ρ. More precisely, given a parameter ρ such that $k = \rho N$, it is natural to ask what is the size of the smallest neighbor set over all ρN-subsets in $[N]$. To the best of our knowledge, this question has only been studied in the context of expander graphs when G is d-regular with $M = N$ and $D = d$ by bounding the second eigenvalue. In [Kah95], Kahale proved that the second eigenvalue can be used to show the graph G is a $(\leq \rho N, \frac{D}{2})$-expander for a $\rho \ll poly(\frac{1}{D})$. Moreover, Kahale showed that some Ramanujan graphs do no expand by more than $D/2$ among small subsets, which indicates $D/2$ is the best parameter for expanders using the eigenvalue method. In another work [WZ99], Wigderson and Zuckerman pointed

out that the expander mixing lemma only helps us determine whether the bipartite graph G is a $(\rho N, (1-\frac{4}{D\rho})N)$-disperser or not, which is not helpful if $D\rho \leq 4$. Even if $D\rho = \Omega(1)$, the expander mixing lemma is unsatisfactory because a random bipartite graph on $[N] \cup [M]$ with right degree d is an $(N, (1-O((1-\rho)^d))M)$-disperser with high probability. Therefore Wigderson and Zuckerman provided an explicit construction for the case $d\rho = \Omega(1)$ when $d = N^{1-\delta+o(1)}$ and $\rho = N^{-(1-\delta)}$ for any $\delta \in (0,1)$. However, there exist graphs such that the second eigenvalue is close to 1 but the graph has very good expansion property among small subsets [KV05, BGH+12]. Therefore the study of the eigenvalue is not enough to fully characterize the vertex expansion. On the other hand, it is well known that a random regular bipartite graph is a good disperser and a good expander simultaneously, it is therefore natural to ask how to certify a random bipartite graph is a good disperser or a good expander.

Our main results are strong integrality gaps and an approximation algorithm for the vertex expansion problem in bipartite graphs. We prove the integrality gaps in the Lasserre hierarchy, which is a strong algorithmic tool in approximation algorithm design such that most currently known semidefinite programming based algorithms can be derived by a constant number of levels in this hierarchy.

We first provide integrality gaps for dispersers in the Lasserre hierarchy. It is well known that a random bipartite graph on $[N] \cup [M]$ is an $(N^\alpha, (1-\delta)M)$-disperser with very high probability when N is large enough and left degree $D = \Theta_{\alpha,\delta}(\log N)$, and these dispersers have wide applications in theoretical computer science [Sha02, Zuc07]. We show an average-case complexity of the disperser problem that given a random bipartite graph, the Lasserre hierarchy cannot approximate the size of the subset in $[N]$ (equivalently the min-entropy of the disperser) required to hit at least 0.01 fraction of vertices in $[M]$ as its neighbors. The second result is an integrality gap for any constant $\rho > 0$ and random bipartite graphs with constant right degree d (the formal statements are in section 10.2.1).

Theorem 10.0.3. *(Informal Statement) For any $\alpha \in (0,1)$ and any $\delta \in (0,1)$, the $N^{\Omega(1)}$-*

level Lasserre hierarchy cannot distinguish whether, for a random bipartite graph G on $[N] \cup [M]$ with left degree $D = O(\log N)$:

1. *G is an $(N^{\alpha}, (1-\delta)M)$-disperser,*
2. *G is not an $(N^{1-\alpha}, \delta M)$-disperser.*

Theorem 10.0.4. *(Informal Statement) For any $\rho > 0$, there exist infinitely many d such that the $\Omega(N)$-level Lasserre hierarchy cannot distinguish whether, for a random bipartite graph G on $[N] \cup [M]$ with right degree d:*

1. *G is a $(\rho N, (1-(1-\rho)^d)M)$-disperser,*
2. *G is not a $\big(\rho N, (1 - C_0 \cdot \frac{1-\rho}{\rho d + 1 - \rho})M\big)$-disperser for a universal constant $C_0 > 0.1$.*

We also provide an approximation algorithm to find a subset of size exact ρN with a relatively small neighbor set when the graph is not a good disperser. For a balanced constant ρ like $\rho \in [1/3, 2/3]$, $\frac{\rho}{1-\rho}$ and $\frac{1-\rho}{\rho}$ are just constants, and the approximation ratio of our algorithm is close to the integrality gap in Theorem 10.0.4 within an extra loss of $\log d$.

Theorem 10.0.5. *Given a bipartite graph $([N], [M])$ that is not a $(\rho N, (1-\Delta)M)$-disperser with right degree d, there exists a polynomial time algorithm that returns a ρN-subset in $[N]$ with a neighbor set of size at most $\big(1 - \Omega(\frac{\min\{(\frac{\rho}{1-\rho})^2, 1\}}{\log d} \cdot d(1-\rho)^d)\Delta\big)M$.*

For expanders, we will show that for any constant $\epsilon > 0$, there is another constant $\epsilon' < \epsilon$ such that the Lasserre hierarchy cannot distinguish the bipartite graph is a $(\rho N, (1-\epsilon')D)$ expander or not a $(\rho N, (1-\epsilon)D)$ expander for small ρ (the formal statement is in Section 10.2.2). To the best knowledge, this is the first hardness result for such an expansion property. For example, it indicates that the Lasserre hierarchy cannot distinguish between a $(\rho N, 0.6322D)$-expander or not a $(\rho N, 0.499D)$-expander.

Theorem 10.0.6. *For any $\epsilon > 0$ and $\epsilon' < \frac{e^{-2\epsilon}-(1-2\epsilon)}{2\epsilon}$, there exist constants ρ and D such that the $\Omega(N)$-level Lasserre hierarchy cannot distinguish whether, a bipartite graph G on $[N] \cup [M]$ with left degree D:*

1. *G is an $(\rho N, (1-\epsilon')D)$-expander,*
2. *G is not an $(\rho N, (1-\epsilon)D)$-expander.*

10.1 Preliminaries

In this chapter, we always use Λ to denote a Constraint Satisfaction Problem and Φ to denote an instance of the CSP. A CSP Λ is specified by a width d, a finite field F_q for a prime power q and a predicate $C \subset F_q^d$. An instance Φ of Λ consists of n variables and m constraints such that every constraint j is in the form of $x_{j,1} \cdots x_{j,d} \in C + \vec{b}_j$ for some $\vec{b}_j \in F_q^d$ and d variables $x_{j,1}, \cdots, x_{j,d}$.

We provide a short description the semidefinite programming relaxations from the Lasserre hierarchy [Las02] (see [Rot13, Bar14] for a complete introduction of Lasserre hierarchy and sum of squares proofs). We will use $f \in \{0,1\}^S$ for $S \subset [N]$ to denote an assignment on variables in $\{x_i | i \in S\}$. Conversely, let f_S denote the partial assignment on S for $f \in \{0,1\}^n$ and $S \subset [n]$. For two assignments $f \in \{0,1\}^S$ and $g \in \{0,1\}^T$ we use $f \circ g$ to denote the assignment on $S \cup T$ when f and g agree on $S \cap T$. For a matrix A, we will use $A_{(i,j)}$ to describe the entry (i,j) of A and $A \succeq 0$ to denote that A is positive semidefinite.

Consider a $\{0,1\}$-programming with an objective function Q and constraints $P_0, \cdots, P_m$,

where $Q, P_0, \cdots, P_m$ are from $\binom{[n]}{\leq d} \times \{0,1\}^d$ to $\mathbb{R}$:

$$\begin{aligned} \max \quad & \sum_{R \in \binom{[n]}{\leq d}, h \in \{0,1\}^R} Q(R,h) 1_{x_R = h} \\ \text{Subject to} \quad & \sum_{R \in \binom{[n]}{\leq d}, h \in \{0,1\}^R} P_j(R,h) 1_{x_R = h} \geq 0 \qquad \forall j \in [m] \\ & x_i \in \{0,1\} \qquad \forall i \in [n] \end{aligned}$$

Let $y_S(f)$ denote the probability that the assignment on S is f in the pseudo-distribution. This $\{0,1\}$-programming[Las02] in the t-level Lasserre hierarchy is:

$$\begin{aligned} \max \quad & \sum_{R \in \binom{[n]}{\leq d}, h \in \{0,1\}^R} Q(R,h) y_R(h) \\ \text{Subject to} \quad & \left(y_{S \cup T}(f \circ g)\right)_{(S \subset \binom{[n]}{\leq t}, f \in \{0,1\}^S), (T \subset \binom{[n]}{\leq t}, g \in \{0,1\}^T)} \succeq 0, \qquad (10.1) \\ & \Big(\sum_{R \in \binom{[n]}{\leq d}, h \in \{0,1\}^R} P_j(R,h) y_{S \cup T \cup R}(f \circ g \circ h) \Big)_{(S \in \binom{[n]}{\leq t}, f \in \{0,1\}^S), (T \in \binom{[n]}{\leq t}, g \in \{0,1\}^T)} \succeq 0, \forall j \in [m]. \\ & \qquad (10.2) \end{aligned}$$

An important tool in the Lasserre hierarchy to prove that the matrices in (10.2) are positive semidefinite is introduced by Guruswami, Sinop and Zhou in [GSZ14], we restate it here and prove it for completeness. Let $u_S(f)$ for all $S \in \binom{[n]}{\leq t}, f \in \{0,1\}^S$ be the vectors to explain the matrix (10.1).

Lemma 10.1.1. *(Restatement of Theorem 2.2 in [GSZ14]) If $\sum_{R,h} P(R,h)\vec{u}_R(h) = \vec{0}$, then the corresponding matrix in (10.2) is positive semidefinite.*

Proof. From the definition of $\vec{u}$, we have

$$\begin{aligned} \sum_{R,h} P(R,h) y_{S \cup T \cup R}(f \circ g \circ h) &= \langle \sum_{R,h} P(R,h) \vec{u}_{S \cup R}(f \circ h), \vec{u}_T(g) \rangle \\ &= \langle \vec{u}_{S \cup T}(f \circ g), \sum_{R,h} P(R,h) \vec{u}_R(h) \rangle = 0. \end{aligned}$$

□

10.1.1 Pairwise Independent Subspaces

We introduce an extra property of pairwise independent subspaces for our construction of integrality gaps of list Constraint Satisfaction Problems.

Definition 10.1.2. *Let C be a pairwise independent subspace of F_q^d and Q be a subset of F_q with size k. We say that C stays in Q with probability p if $\Pr_{x\sim C}[x \in Q^d] = \frac{|C\cap Q^d|}{|C|} \geq p$.*

In [BGGP12], Benjamini et.al. proved $p \leq \frac{k/q}{(1-k/q)\cdot d+k/q}$ for infinitely many d when $|Q| = k$. They also provided a distribution that matches the upper bound with probability $\frac{k/q}{(1-k/q)\cdot d+k/q}$ for every d with $q|(d-1)$. In this thesis, we need the property that C is a subspace rather than an arbitrary distribution in F_q^d. We provide two constructions for the base cases $k = 1$ and $k = q-1$.

Lemma 10.1.3. *There exist infinitely many d such that there is a pairwise independent subspace $C \subset F_q^d$ that stays in a size 1 subset Q of F_q with probability $\frac{1/q}{(1-1/q)d+1/q}$.*

Proof. Choose $Q = \{0\}$ and C to be the dual code of Hamming codes over F_q with block length $d = \frac{q^l-1}{q-1}$ and distance 3 for an integer l. Using $|C| = q^l$, the probability is $\frac{1}{|C|} = \frac{1/q}{(1-1/q)d+1/q}$. It is pairwise independent because the dual distance of C is 3. □

Lemma 10.1.4. *There exist infinitely many d such that there is a pairwise independent subspace $C \subset F_q^d$ staying in a $(q-1)$-subset Q of F_q with probability at least $\Omega(\frac{(q-1)/q}{d/q+(q-1)/q})$.*

Proof. First, we provide a construction for $d = q-1$ then generalize it to $d = (q-1)q^l$ for any integer l. For $d = q-1$, the generator matrix of the subspace is a $(q-1)\times 2$ matrix where row i is (α_i, α_i^2) for $q-1$ distinct elements $\{\alpha_1, \cdots, \alpha_{q-1}\} = F_q^*$. Because $\alpha_i \neq \alpha_j$ for any two different rows i and j, it is pairwise independent. Let $Q = F_q \setminus \{1\}$. Using the inclusion-exclusion principle and the fact that a quadratic equation can have at most 2 roots

in F_q:

$$\begin{aligned}
\Pr_{x\sim C}[x \in Q^d] &= \Pr_{x\sim C}[\forall \beta \in F_q^*, x_\beta \neq 1] \\
&= 1 - \sum_{\beta \in F_q^*} Pr[x_\beta = 1] + \sum_{\{\beta_1,\beta_2\}\in\binom{F_q^*}{2}} Pr[x_{\beta_1} = 1 \wedge x_{\beta_2} = 1] \\
&- \sum_{\{\beta_1,\beta_2,\beta_3\}\in\binom{F_q^*}{3}} Pr[x_{\beta_1} = 1 \wedge x_{\beta_2} = 1 \wedge x_{\beta_3} = 1] + \cdots \\
&= 1 - \frac{q-1}{q} + \frac{\binom{q-1}{2}}{q^2} - 0 + 0 \\
&= \frac{q^2-q+2}{2q^2} \geq \frac{1}{2} - \frac{1}{2q}
\end{aligned}$$

For any $d = (q-1)q^l$, the generator matrix of the subspace is a $d \times (l+2)$ matrix where every row is in the form $(\alpha, \alpha^2, \beta_1, \cdots, \beta_l)$ for all nonzero elements $\alpha \in F_q^*$ and $\beta_1 \in F_q, \cdots, \beta_l \in F_q$. The pairwise independence comes from a similar analysis. $\Pr_{x\sim C}[x \in Q^d] \geq \frac{1}{q^l}(\frac{1}{2} - \frac{1}{2q})$ because it is as same as $d = q-1$ when all coefficients before $\beta_1, \cdots, \beta_l$ are 0, which is $\geq \frac{1}{3} \cdot \frac{q-1}{d+q-1}$. $\square$

Remark 10.1.5. *The construction for $d = q-1$ also provides a subspace that stays in Q with probability $1 - \frac{d}{q} + \frac{\binom{d}{2}}{q^2}$ for any $d < q-1$ by deleting unnecessary rows in the matrix.*

10.2 Integrality Gaps

We first consider a natural $\{0,1\}$ programming to determine the vertex expansion of ρN-subsets in $[N]$ given a bipartite graph $G = ([N], [M], E)$:

$$\begin{aligned}
&\min \sum_{j=1}^{M} \vee_{i\in\Gamma(j)} x_i = \min \sum_{j=1}^{M} (1 - 1_{\forall i\in\Gamma(j), x_i = 0}) \\
\text{Subject to } &\sum_{i=1}^{N} x_i \geq \rho N \\
&x_i \in \{0,1\} \qquad\qquad \text{for every } i \in [N]
\end{aligned}$$

We relax it to a convex programming in the t-level Lasserre hierarchy.

$$\min \sum_{j=1}^{M}(1 - y_{\Gamma(j)}(\vec{0}))$$

$$\text{Subject to } \big(y_{S\cup T}(f \circ g)\big)_{((S\in\binom{[N]}{\leq t}, f\in\{0,1\}^S),(T\in\binom{[N]}{\leq t}, g\in\{0,1\}^T))} \succeq 0 \tag{10.3}$$

$$\big(\sum_{i=1}^{N} y_{S\cup T\cup\{i\}}(f \circ g \ \circ 1) - \rho N \cdot y_{S\cup T}(f \circ g)\big)_{((S\in\binom{[N]}{\leq t}, f\in\{0,1\}^S),(T\in\binom{[N]}{\leq t}, g\in\{0,1\}^T))} \succeq 0 \tag{10.4}$$

In this section, we focus on random bipartite graphs G on $[N]\cup[M]$ that are d-regular in $[M]$, which are generated by connecting each vertex in $[M]$ to d random vertices in $[N]$ independently. The main technical result we will prove in this section is:

Lemma 10.2.1. *Suppose there is a pairwise independent subspace $C \subseteq F_q^d$ staying in a k-subset with probability $\geq p_0$. Let $G = ([N], [M], E)$ be a random bipartite graph with $M = O(N)$ that is d-regular in $[M]$, the $\Omega(N)$-level Lasserre hierarchy for G and $\rho = 1 - k/q$ has an objective value at most $(1 - p_0 + \frac{1}{N^{1/3}})M$ with high probability.*

We introduce list Constraint Satisfaction Problems which allow every variable to take k values from the alphabet. Next, we lower bound the objective value of an instance of a list CSP in the Lasserre hierarchy from the objective value of the corresponding instance of the CSP in the Lassrre hierarchy. Then we show how to use list CSPs to obtain an upper bound of the vertex expansion for $\rho = 1 - k/q$ in the Lasserre hierarchy.

Definition 10.2.2 (list Constraint Satisfaction Problem)**.** *A list Constraint Satisfaction Problem (list CSP) Λ is specified by a constant k, a width d, a domain over finite field F_q for a prime power q, and a predicate $C \subseteq F_q^d$. An instance Φ of Λ consists of a set of variables $\{x_1, \cdots, x_n\}$ and a set of constraints $\{C_1, C_2, \cdots, C_m\}$ on the variables. Every variable x_i takes k values in F_q, and every constraint C_j consists of a set of d variables $x_{j,1}, x_{j,2}, \cdots, x_{j,d}$*

and an assignment $\vec{b_j} \in F_q^d$. *The value of* C_j *is* $|(C+\vec{b_j}) \cap x_{i,1} \times x_{i,2} \cdots x_{i,d}| \in \mathbb{N}$. *The value of* Φ *is the summation of values over all constraints, and the objective is to find an assignment on* $\{x_1, \cdots, x_n\}$ *that maximizes the total value as large as possible.*

Remark 10.2.3. *We abuse the notation* C_j *to denote the variable subset* $\{x_{j,1}, x_{j,2}, \cdots, x_{j,d}\}$. *Our definition is consistent with the definition of the classical CSP when* $k = 1$. *The differences between a list CSP and a classical CSP are that a list CSP allow each variable to choose* k *values in* F_q *instead of one value and relax every constraint* C_i *from* $F_q^d \to \{0,1\}$ *to* $F_q^d \to \mathbb{N}$.

The $\{0,1\}$ programming for an instance Φ with variables $\{x_1, \cdots, x_n\}$ and constraints $\{C_1, \cdots, C_m\}$ of Λ with parameters k, F_q,and a predicate C states as follows (the variable set is the direct product of $[n]$ and F_q in the $\{0,1\}$ programming):

$$\begin{aligned} \max \sum_{j\in[m]} \sum_{f\in C+\vec{b}_j} & 1_{\forall i\in C(j), x_{i,f(i)}=1} & \\ \text{Subject to } x_{i,\alpha} &\in \{0,1\} & \forall (i,\alpha) \in [n] \times F_q \\ \sum_{\alpha\in F_q} x_{i,\alpha} &= k & \forall i \in [n] \end{aligned}$$

The SDP in the t-level Lasserre hierarchy for Φ succeeds this $\{0,1\}$ programming as follows:

$$\max \sum_{j\in[m]} \sum_{f\in C+\vec{b}_j} y_{(C_j,f)}(\vec{1})$$

$$\text{S.t.} \big(y_{S\cup T}(f\circ g)\big)_{(S\subset \binom{[n]\times F_q}{\leq t}, f\in\{0,1\}^S), (T\subset \binom{[n]\times F_q}{\leq t}, g\in\{0,1\}^T)} \succeq 0 \tag{10.5}$$

$$\big(k\cdot y_{S\cup T}(f\circ g) - \sum_{\alpha} y_{S\cup T\cup\{(i,\alpha)\}}(f\circ g\circ 1)\big)_{(S\subset \binom{[n]\times F_q}{\leq t}, f\in\{0,1\}^S), (T\subset \binom{[n]\times F_q}{\leq t}, g\in\{0,1\}^T)} = 0, \forall i \in [n] \tag{10.6}$$

Definition 10.2.4. *Let* Λ *be the list CSP problem with parameters* k, q, d *and a predicate* $C \subset F_q^d$. *Let* Φ *be an instance of* Λ *with* n *variables and* m *constraints.* $p(\Phi)$ *is the projection*

instance from Φ *in the CSP of the same parameters* $q, d, C \subseteq F_q^d$*, and the same constraints* $(C_1, \vec{b}_1), (C_2, \vec{b}_2), \cdots, (C_m, \vec{b}_m)$ *except* $k = 1$*.*

Recall that a subspace $C \subset F_q^d$ stays in a subset $Q \subset F_q$ with probability p_0 if $\Pr_{x \sim C}[x \in Q^d] \geq p_0$. We lower bound Φ's objective value in the Lasserre hierarchy by exploiting the subspace property of C and Q.

Lemma 10.2.5. *Let* Φ *be an instance of the list CSP* Λ *with parameters* k, q, d *and a predicate* C*, where* C *is a subspace of* F_q^d *staying in a* k*-subset* Q *with probability at least* p_0*. Suppose* $p(\Phi)$*'s value is* γ *in the* w*-level Lasserre hierarchy, then* Φ*'s value is at least* $p_0|C| \cdot \gamma$ *in the* w*-level Lasserre hierarchy.*

Proof. Let $y_S(f)$ and $\vec{v}_S(f)$ for $S \in \binom{[n] \times F_q}{\leq w}$ and $f \in \{0,1\}^S$ denote the pseudo-distribution and the vectors in the w-level Lasserre hierarchy for $p(\Phi)$ respectively. Let z and $\vec{u}$ denote the pseudo-distribution and vectors in the w-level Lasserre hierarchy for Φ. The construction of z and $\vec{u}$ from y and $\vec{v}$ are based on the subspace C and Q. The intuition is to choose $x_i = \alpha + Q$ in Φ if $x_i = \alpha$ for some $\alpha \in F_q$ in $p(\Phi)$.

Before constructing z and $\vec{u}$, define $\oplus$ operation as follows. For any $S \in \binom{[n] \times F_q}{\leq w}, g \in \{0,1\}^S$, and $P \subseteq F_q$, let $S \oplus P$ denote the union of the subset $(i, \alpha + P)$ for every element (i, α) in S, which is $\cup_{(i,\alpha) \in S}\{(i, \alpha + P)\}$ in $[n] \times F_q$, and $g \oplus P \in \{0,1\}^{S \oplus P}$ denote the assignment on $S \oplus P$ such that $g \oplus P(i, \alpha + P) = g(i, \alpha)$. If there is a conflict in the definition of $g \oplus P$, namely $\exists (i, \beta)$ such that $(i, \beta) \in (i, \alpha_1 + P)$ and $(i, \beta) \in (i, \alpha_2 + P)$ for two distinct $(i, \alpha_1), (i, \alpha_2)$ in S, define $g \oplus P(i, \beta)$ to be an arbitrary one. Because every variable only takes one value in $p(\Phi)$, $y_S(g) = 0$ if there is a conflict on $g \oplus P \in \{0,1\}^{S \oplus P}$. Follow the

intuition mentioned above, for any $S \subset \binom{[n]\times F_q}{\leq w}$ and $g \in \{0,1\}^S$, let $R = \{i | \exists \alpha, (i,\alpha) \in S\}$,

$$z_S(g) = \sum_{T \in \binom{R\times F_q}{\leq w}, g' \in \{0,1\}^T : S \subseteq T \oplus Q, g' \oplus Q(S) = g} y_T(g'),$$

$$\vec{u}_S(g) = \sum_{T \in \binom{R\times F_q}{\leq w}, g' \in \{0,1\}^T : S \subseteq T \oplus Q, g' \oplus Q(S) = g} \vec{v}_T(g').$$

The verification of the fact that $\vec{u}$ explains z in (10.5) of Φ is straightforward. To verify (10.6) is positive semidefinite, notice that every variable x_i takes k values in F_q:

$$\begin{aligned}\sum_{\alpha \in F_q} z_{(i,\alpha)}(1) &= \sum_{\alpha \in F_q} \sum_{\beta \in Q} y_{(i,\alpha-\beta)}(1) \\ &= \sum_{\beta \in Q} \sum_{\alpha \in F_q} y_{(i,\alpha-\beta)}(1) = |Q| = k.\end{aligned}$$

By a similar analysis, $\sum_{\alpha \in F_q} \vec{u}_{(i,\alpha)}(1) = k\vec{v}_\emptyset$ and apply Lemma 10.1.1 to prove (10.6) is PSD.

Recall that $p(\Phi)$'s value is $\sum_{j\in[m]} \sum_{f \in C + \vec{b}_j} y_{(C_j,f)}(\vec{1}) = \gamma$, so Φ's objective value in the w-level Lasserre hierarchy is

$$\begin{aligned}\sum_{j\in[m]} \sum_{f \in C+\vec{b}_j} z_{(C_j,f)}(\vec{1}) &= \sum_{j\in[m]} \sum_{f \in C+\vec{b}_j} \sum_{f' \in F_q^d : f \in f' \oplus Q} y_{(C_j,f')}(\vec{1}) \\ &= \sum_{j\in[m]} \sum_{f' \in F_q^d} \sum_{f \in C+\vec{b}_j} y_{(C_j,f')}(\vec{1}) \cdot 1_{f \in f' \oplus Q} \\ &\geq \sum_{j\in[m]} \sum_{f' \in C+\vec{b}_j} y_{(C_j,f')}(\vec{1}) \cdot |(f' \oplus Q) \cap (C + \vec{b}_j)| \\ &\geq \sum_{j\in[m]} \sum_{f' \in C+\vec{b}_j} y_{(C_j,f')}(\vec{1}) \cdot p_0|C| \\ &\geq p_0|C| \cdot \gamma.\end{aligned}$$

□

Before proving Lemma 10.2.1, We restate Theorem G.8 that is summarized by Chan in [Cha13] of the previous works [Gri01a, Sch08, Tul09] and observe that the pseudo-distirbution in their construction is uniform over C on every constraint.

Theorem 10.2.6. *([Cha13]) Let F_q be the finite field with size q and C be a pairwise independent subspace of F_q^d for some constant $d \geq 3$. The CSP is specified by parameters $F_q, d, k = 1$ and a predicate C. The value of an instance Φ of this CSP on n variables with m constraints is m in the $\Omega(t)$-level Lasserre hierarchy if every subset T of at most t constraints contains at least $(d - 1.4)|T|$ variables.*

Observation 10.2.7. *Let $y_S(\{0,1\}^S)$ denote the pseudo-distribution on S provided by the solution of the semidefinite programming in the Lasserre hierarchy of Φ. For every constraint $C_j (j \in [m])$ in Φ, $y_{C_j}(\{0,1\}^{C_j})$ provides a uniform distribution over all assignments that satisfy constraint C_j.*

Proof of Lemma 10.2.1. Without lose of generality, we assume $[N] = [n] \times F_q$. It is natural to think $[N]$ corresponding to n variables and each variables has q vertices corresponding to F_q. Let G be a random bipartite graph on $[N] \cup [M]$ that is d-regular on $[M]$. For each vertex $j \in M$, the probability that j has two or more neighbors in $i \times F_q$ for some i is at most $\frac{d^2 q}{n}$. Let R denote the subset in M that do not have two or more neighbors in any $i \times F_q$ for all $i \in [n]$. With probability at least $1 - \frac{1}{\sqrt{n}}$, $R \geq (1 - \frac{d^2 q}{\sqrt{n}})M$.

Because the neighbors of each vertex in $[M]$ is generated by choosing d random vertices in $[N]$. For each vertex in R, the generation of its neighbors is as same as first sampling d random variables in $[n]$ then sampling an element in F_q for each variable. By a standard calculation using Chernoff bound and Stirling formula, there exists a constant $\beta = O_{d,M/n}(1)$ such that with high probability, $\forall T \subseteq \binom{R}{\leq \beta n}$, T contains at least $(d - 1.4)|T|$ variables.

We construct an instance Φ based on the induced graph of $[n] \times F_q \cup R$ in the list CSP with the parameters k, q, d and the predicate $\{\vec{0}\}$. For each vertex $j \in R$, let $(i_1, b_1), \cdots, (i_d, b_d)$ be its neighbors in G. We add a constraint C_j in Φ with variables $x_{i_1}, \cdots, x_{i_d}$ and $\vec{b} = (b_1, \cdots, b_d)$.

Recall that C is a subspace staying a subset Q of size k with probability p_0, we use the following two claims to prove the value of the vertex expansion of ρN-subsets in the Lasserre hierarchy is at most $(1-p_0)R+(M-R) \leq (1-p_0)(1-\frac{d^2q}{\sqrt{n}})M + \frac{d^2q}{\sqrt{n}}M \leq (1-p_0+o(1))M$ with high probability.

Claim 10.2.8. *Φ's value is at least $p_0|R|$ in the $\Omega(\beta n)$-level Lasserre hierarchy.*

Claim 10.2.9. *Suppose Φ's value is at least r in the t-level Lasserre hierarchy, the objective value of the t-level Lasserre hierarchy is at most $|R|-r$ for the vertex expansion problem on $[N] \cup R$ with $\rho = 1-k/q$.*

□

Proof of Claim 10.2.8. Let Λ be the list CSP with parameters F_q, k, d and predicate C. Let Φ' be the instance of Φ in Λ. From Theorem 10.2.6, $P(\Phi')$'s value is R because every small constraint subset contains at least $(d-1.4)|T|$ variables. From Lemma 10.2.5, Φ's value is at least $p_0|C| \cdot R$ in the $\Omega(n)$-level Lasserre hierarchy.

Let us take a closer look, for each constraint j in $P(\Phi')$, the pseudo-distribution on C_j is uniformly distributed over $b_j + C$. Therefore every assignment $f + b_j$ for $f \in C$ appears in the pseudo-distribution of $P(\Phi')$ on C_j with probability $1/|C|$. As the same reason, every assignment $f + b_j$ appears in the pseudo-distribution of Φ' with the same probability $\frac{|Q^d \cap C|}{C} = p_0$. Because $\vec{0} \in C$, the probability C_j contains $\vec{0} + \vec{b}_j$ in the pseudo-distribution of Φ' is p_0 by the analysis. Using the solution of Φ' in the $\Omega(\beta n)$-level Lasserre hierarchy as the solution of Φ, it is easy to see Φ's value is at least $p_0|R|$. □

Proof of Claim 10.2.9. Let $y_S(f), \vec{v}_S(f)$ for all $S \subseteq \binom{[n] \times F_q}{t}$ and $f \in \{0,1\}^S$ be the solution of pseudodistribution and vectors in the t-level Lasserre hierarchy for Φ. We define

$z_S(f), \vec{u}_S(f)$ for all $S \subseteq \binom{[n]\times F_q}{t}([N] = [n] \times F_q)$ and $f \in \{0,1\}^S$ to be the pseudodistribution and vectors for the vertex expansion problem as follows:

$$\vec{u}_S(f) = \vec{v}_S(\vec{1} - f), z_S(f) = y_S(\vec{1} - f).$$

The verification of the fact that $\vec{u}$ explains the matrix (10.3) of z in the Lasserre hierarchy is straightforward. Another property from the construction is

$$\begin{aligned}\sum_{(x_i,b)} \vec{u}_{(x_i,b)}(1) = \sum_{(x_i,b)} \vec{v}_{(x_i,b)}(0) &= \sum_{(x_i,b)} (\vec{v}_\emptyset - \vec{v}_{(x_i,b)}(1)) \\ &= \sum_{i\in[n]} \sum_{b\ \in F_q} (\vec{v}_\emptyset - \vec{v}_{(x_i,b)}(1)) = \sum_i (q\vec{v}_\emptyset - k\vec{v}_\emptyset) = \rho N \cdot \vec{v}_\emptyset,\end{aligned}$$

which implies the matrix in (10.4) is positive semidefinite by Lemma 10.1.1.

The value of the vertex expansion problem given $z, \vec{u}$ is $\sum_{j\in[R]}(1 - z_{N(j)}(\vec{0})) = \sum_{j\in[R]}(1 - y_{(N(j))}(\vec{1})) = R - \sum_{j\in[R]} y_{(N(j))}(\vec{1}) = R - r$. □

On the other hand, it is easy to prove a random bipartite graph has very good vertex expansion by using Chernoff bound and union bound.

Lemma 10.2.10. *For any constants $d, \rho, \epsilon > 0$, and $c \geq \frac{20q}{(1-\rho)^d \cdot \epsilon^2}$, with high probability, a random bipartite graph on $[N] \cup [M] (M = cN)$ that is d-regular in $[M]$ guarantees that every ρN-subset in $[N]$ contains at least $1 - (1+\epsilon)(1-\rho)^d$ different vertices in $[M]$.*

Proof. For any subset $S \subseteq [N]$ of size ρN, the probability that a vertex in $[M]$ is not a neighbor of S is at most $(1-\rho)^d + o(1)$. Applying Chernoff bound on M independent experiments, the probability that S contains less than $(1 - (1+\epsilon)(1-\rho)^d)$ neighbors in $[M]$ is at most $exp(-\epsilon^2(1-\rho)^d M/12) \leq 2^{-M}$. From union bound, every ρN subset has at least $(1 - (1+\epsilon)(1-\rho)^d)$ neighbors with high probability. □

10.2.1 Integrality Gaps for Dispersers

Theorem 10.2.11. *For any $\epsilon > 0$ and $\rho \in (0,1)$, there exist infinitely many d such that a random bipartite graph on $[N] \cup [M]$ that is d-regular in $[M]$ satisfies the following two properties with high probability:*

1. *It is a $\big(\rho N, (1-(1-\rho)^d - \epsilon)M\big)$-disperser.*
2. *The objective value of the $\Omega(N)$-level Lasserre hierarchy for ρ is at most $\big(1 - C_0 \cdot \frac{1-\rho}{d\rho+1-\rho}\big)M$ for a universal constant $C_0 \geq 1/10$.*

Proof. Let $M \geq \frac{20q}{(1-\rho)^d \cdot \epsilon^2} N$, a random bipartite graph G is a $\big(\rho N, (1-(1-\rho)^d - \epsilon)M\big)$-disperser from Lemma 10.2.10 with very high probability.

On the other hand, choose a prime power q and k in the base cases of Lemma 10.1.3 or Lemma 10.1.4 such that $\rho' = 1 - k/q > \rho$ and p_0 be the probability that the subspace C staying in a k-subset. From the construction, $p_0 \geq \frac{1}{3}\frac{1-\rho'}{d\rho'+1-\rho'} \geq \frac{1}{9} \cdot \frac{1-\rho}{d\rho+1-\rho}$. From Lemma 10.2.1, a random graph G that is d-regular in $[M]$ has vertex expansion at most $(1-p_0)M$ for ρ' with high probability. Because $\rho' \geq \rho$, this indicates The objective value of the $\Omega(N)$-level Lasserre hierarchy for ρ is at most $(1 - \frac{1}{9} \cdot \frac{1-\rho}{d\rho+1-\rho})M$.Therefore, a random bipartite graph G satisfies the two properties with high probability. □

We generalize the above construction to $d = \Theta(\log N)$ and prove the Lasserre hierarchy cannot approximate the entropy of a disperser in the rest of this section. Because $d = \Theta(\log N)$ is a super constant, we relax the strong requirement in the variable expansion of constraints and follow the approach of [Tul09]. We also notice the same observation has independently provided by Bhaskara et.al. in [BCV+12].

Theorem 10.2.12. *(Restatement of Theorem 4.3 in [Tul09]) Let C be the dual space of a linear codes with dimension d and distance l over F_q. Let Φ with n variables and m*

constraints be an instance of the CSP Λ *with* $d, k = 1, F_q$ *and predicate* C*. If for every subset* S *of at most* t *constraints in* Φ*, it contains at least* $(1 - l/2 + .2)d \cdot |S|$ *different variables. Then the value of* Φ *is* m *in the* $\Omega(t)$*-level Lasserre hierarchy.*

Lemma 10.2.13. *For any prime power* $q, \epsilon > 0, \delta > 0$*, and any constant* c*, a random bipartite graph on* $[N] \cup [M]$ *that is* $d = c \log N$*-regular in* M *has the following two properties with high probability:*

1. *It is a* $(\delta N, (1 - 2(1 - \delta)^d)M)$*-disperser.*
2. *The objective value of the* $N^{\Omega(1)}$*-level Lasserre hierarchy for* $\rho = \frac{q-1}{q}$ *is at most* $(1 - q^{-\epsilon d} + \frac{1}{N^{1/3}})M$*.*

Proof. Let A be a linear code over F_q with dimension d, rate $(1 - \epsilon)d$ and distance $3\gamma d$ for some $\gamma > 0$. C is the dual space of A with size $|C| = q^{\epsilon d}$. Let $M = \frac{20q \cdot N}{(1-\delta)^d}$, which is $poly(N)$ here. From Lemma 10.2.10, a random bipartite graph G on $[N] \cup [M]$ that is d-regular in M is a $(\delta N, (1 - 2(1 - \delta)^d)M)$-disperser with very high probability.

In the rest of proof, it is enough to show that for every subsets $S \subseteq \binom{[M]}{\leq N^{\gamma/2}}$ in Φ, the constraints in S contain at least $(1 - \gamma)|S|d$ variables. By union bound, the probability that does not happen is bounded by

$$\sum_{l=1}^{N^{\gamma/2}} \binom{M}{l} \binom{N}{(1-\gamma)d \cdot l} (\frac{(1-\gamma)dl}{N})^{dl} \leq \sum_{l \leq N^{\gamma/2}} M^l N^{(1-\gamma)dl} (\frac{dl}{N})^{dl} \leq \sum_{l \leq N^{\gamma/2}} \frac{M^l}{N^{\gamma \cdot dl/2}} \frac{(dl)^{dl}}{N^{\gamma \cdot dl/2}} \leq 0.1.$$

By Lemma 10.2.1, the value of G with $\rho = \frac{q-1}{q}$ is at most $(1 - 1/|C| + \frac{d^2 q}{\sqrt{N}})M \leq (1 - q^{-\epsilon d} + \frac{1}{N^{1/3}})M$ in the $\Omega(n^{\gamma/2})$-level Lasserre hierarchy. □

We show the equivalence between the vertex expansion problem and the problem of approximating the entropy in a disperser:

Problem 10.2.14. *Given a bipartite graph $([N], [M], E)$ and ρ, determine the size of the smallest neighbor set over all subsets of size at least ρN in $[N]$.*

Problem 10.2.15. *Given a bipartite graph $([N], [M], E)$ and γ, determine the size of the largest subset in $[N]$ with a neighbor set of size $\leq \gamma M$.*

We prove the equivalence of these two problems with parameters $\rho + \gamma = 1$. For a bipartite graph $([N], [M], E)$ and a parameter γ, let T be the largest subset in $[N]$ with $|\Gamma(T)| \leq \gamma M$. Let $S = [M] \setminus \Gamma(T)$. Then $|S| \geq (1-\gamma)M$ and $\Gamma(S) \subseteq [N] \setminus T$. Since T is the largest subset with $|\Gamma(T)| \leq \gamma M$, S is the subset of size at least $(1-\gamma)M$ with the smallest neighbor set. The converse is similar, which shows the equivalence between these two problems.

Theorem 10.2.16. *For any $\alpha \in (0,1)$, any $\delta \in (0,1)$ and any prime power q, there exists a constant c such that a random bipartite graph on $[N] \cup [M]$ that is $D = c \log N$-regular in $[N]$ has the following two properties with high probability:*

1. *It is an $(N^\alpha, (1-\delta)M)$-disperser.*
2. *The objective value of the SDP in the $N^{\Omega(1)}$-level Lasserre hierarchy for obtaining M/q distinct neighbors is at least $N^{1-\alpha/2}$.*

Proof. Let $\epsilon = \frac{\log \frac{1}{1-\delta}}{4\alpha \log q} = O(1)$ and $d = \frac{\log N}{4\epsilon \log q}$ such that $|C| = q^{\epsilon d} = N^{1/4}$ and $M = \frac{20q \cdot N}{(1-\delta)^d} \geq N^{1/\alpha}$. So $d = O(\log M)$.

From Lemma 10.2.13, a random bipartite graph on $[N] \cup [M]$ d-regular in $[M]$ is a $(\delta N, M - M^\alpha)$-disperser, but the value of $N^{\Omega(1)}$-level Lasserre hierarchy for G with $\rho = 1 - 1/q$ is at most $M - M^{1-\alpha/2}$. From the equivalence, any subset of size M^α in $[M]$ has a neighbor set of size at least $(1-\delta)N$. On the other hand, it is possible that there exists a $M^{1-\alpha/2}$-subset of $[M]$ with a neighbor set of size at most $[N]/q$ in the Lasserre hierarchy, from the

fact that the $N^{\Omega(1)}$-level Lasserre hierarchy has a value at most $M - M^{1-\alpha/2}$ for $\rho = 1 - 1/q$. To finish the proof, swap $[N]$ and $[M]$ in the bipartite graph such that $D = d$ in the new bipartite graph. □

Corollary 10.2.17. *(Restatement of Theorem 10.0.3) For any $\alpha \in (0,1)$, any $\delta \in (0,1)$, there exists a constant c such that a random bipartite graph on $[N] \cup [M]$ with $D = c \log N$-regular in $[N]$ has the following two properties with high probability:*

1. *It is an $(N^\alpha, (1-\delta)M)$-disperser.*
2. *The objective value of the SDP in the $N^{\Omega(1)}$-level Lasserre hierarchy for obtaining δM distinct neighbors is at least $N^{1-\alpha}$.*

10.2.2 Integrality Gaps for Expanders

We prove that a random bipartite graph is almost D-reguar on the right hand side and use the fact $dN \approx DM$.

Theorem 10.2.18. *For any prime power q, integer $d < q$ and constant $\delta > 0$, there exist a constant D and a bipartite graph G on $[N] \cup [M]$ with the largest left degree D and the largest right degree d has the following two properties for $\rho = 1/q$:*

1. *It is a $(\rho N, (1-\epsilon'-2\delta)D)$-expander with $\epsilon' = \frac{(1-\rho)^d - (1-\rho d)}{\rho d} = \sum_{i=1}^{d-1} (-1)^{i-1} \frac{(d-1)\cdots(d-i+1)}{(i+1)!} \rho^i$.*
2. *The objective value of the vertex expansion for G with ρ in the $\Omega(N)$-level Lasserre hierarchy is at most $(1 - \epsilon + \delta)D \cdot \rho N$ with $\epsilon = \frac{\rho(d-1)}{2}$.*

Proof. Let β be a very small constant specified later and $c = \frac{100q \cdot \log(1/\beta)}{d(1-\rho)^d \cdot \delta^2}$. Let G_0 be a random graph on $[N] \cup [M]$ with $M = cN$ that is d-regular in $[M]$. Let $D_0 = \frac{dM}{N}$ and L denote the vertices in $[N]$ with degree $[(1-\delta)D_0, (1+\delta)D_0]$. Let G_1 denote the induced graph of G_0

on $L \cup [M]$. The largest degree of L is $D = (1+\delta)D_0$ and the largest degree of M is d. We will prove G_1 is a bipartite graph that satisfies the two properties in this lemma with high probability. Because G_0 is a random graph, we assume there exists a constant $\gamma = O_{M/N,d}(1)$ such that every subset $S \in \binom{M}{\leq \gamma N}$ has different $(d-1.1)|S|$ neighbors.

In expectation, each vertex in N has degree D_0. By Chernoff bound, the fraction of vertices in $[N]$ of G_0 with degree more than $(1+\delta)D_0$ or less than $(1-\delta)D_0$ is at most $2exp(-\delta^2 \cdot \frac{d}{N} \cdot M/12) \leq \beta^4$. At the same time, with high probability, G_0 satisfies that any $\beta^3 N$-subset in $[N]$ has total degree at most βdM because $\binom{N}{\beta^3 N} \cdot exp(-(\frac{1}{\beta^2})^2(\beta^3 d) \cdot M/12)$ is exponentially small in N. Therefore with high probability, $|L| \geq (1-\beta^3)N$ and there are at least $(1-\beta)dM$ edges in G_1.

We first verify the objective value of the vertex expansion for G_1 with $\rho = 1/q$ in the $\Omega(N)$-level Lasserre hierarchy is at most $(1-\epsilon+\delta)D \cdot \rho N)$. Let R be the vertices in $[M]$ that have degree d. From Lemma 10.2.1, the objective value of the vertex expansion for $L \cup R$ with $\rho = 1/q$ in the $\Omega(\gamma N)$-level Lasserre hierarchy is at most $(1-p_0)|R|$ where p_0 is the staying probability of C in a $q-1$ subset. From Lemma 10.1.4, $p_0 = 1 - d\rho + \binom{d}{2}\rho^2$. Therefore $(1-p_0)|R| \geq (1-1+d\rho-\binom{d}{2}\rho^2)(1-d\beta)M$. For the vertices in $M \setminus R$, they will contribute at most $d\beta M$ in the objective value of the Lasserre hierarchy. Therefore the objective value for G_1 is at most $(d\rho - \binom{d}{2}\rho^2 + d\beta)M = (1 - \frac{(d-1)\rho}{2} + \frac{\beta}{\rho})\rho dM \leq (1-\epsilon+\frac{\beta}{\rho})\rho DM$.

For the integral value, every ρN-subset in $[N]$ has at least $(1-(1+\beta)(1-\rho)^d)M$ neighbors in G_0 by Lemma 10.2.10. Because G_1 is the induced graph of G_0 on $L \cup [M]$, every ρN-subset in L has at least $(1-(1+\beta)(1-\rho)^d)M \geq (1-\epsilon'-\frac{\beta}{\rho d})\rho dM \geq (1-\epsilon'-\frac{\beta}{\rho d})D_0 \cdot \rho N$ neighbors in G_1. By setting β small enough, there exists a bipartite graph with the required two properties. □

Corollary 10.2.19. *For any $\epsilon > 0$ and any $\epsilon' < \frac{e^{-2\epsilon}-(1-2\epsilon)}{2\epsilon}$, there exist ρ small enough and a bipartite graph G with the largest left degree $D = O(1)$ that has the following two properties:*

1. *It is a* $(\rho N, (1-\epsilon')D)$*-expander.*

2. *The objective value of the vertex expansion for* G *with* ρ *in the* $\Omega(N)$*-level Lasserre hierarchy is at most* $(1-\epsilon)D \cdot \rho N$*.*

Proof. Think ρ to be a small constant and $d = \frac{2\epsilon}{\rho} + 1$ such that ϵ is very close to $\frac{\rho d}{2}$. Then the limit of $\epsilon' = \frac{(1-\rho)^d - (1-\rho d)}{\rho d}$ is $\frac{e^{-\rho d} - (1-\rho d)}{\rho d} = \frac{e^{-2\epsilon} - (1-2\epsilon)}{2\epsilon}$ by decreasing ρ. □

10.3 An Approximation algorithm for Dispersers

In this section, we will provide a polynomial time algorithm that has an approximation ratio close to the integrality gap in Theorem 10.0.4.

Theorem 10.3.1. *Given a bipartite graph* $([N], [M], E)$ *with right degree* d*, if* $(1-\Delta)M$ *is the size of the smallest neighbor set over* ρN*-subsets in* $[N]$*, there exists a polynomial time algorithm that outputs a subset* $T \subseteq [N]$*, such that* $|T| = \rho N$ *and* $\Gamma(T) \leq \big(1 - \Omega(\frac{\min\{(\frac{\rho}{1-\rho})^2, 1\}}{\log d} \cdot d(1-\rho)^d \cdot \Delta)\big)M$*.*

We consider a simple semidefinite programming for finding a subset $T \subseteq [N]$ that maximizes the number of unconnected vertices to T.

$$\begin{aligned} & \max \sum_{j \in [M]} \| \frac{1}{d} \sum_{i \in \Gamma(j)} \vec{v}_i \|_2^2 && (*) \\ \text{Subject to} \quad & \langle \vec{v}_i, \vec{v}_i \rangle \leq 1 \\ & \sum_{i=1}^{n} \vec{v}_i = \vec{0} \end{aligned}$$

We first show the objective value of the semidefinite programming is at least $\min\{(\frac{\rho}{1-\rho})^2, 1\} \cdot \Delta$. For convenience, let δ denote the value of this semidefinite programming and A denote the positive definite matrix of the objective function in the semidefinite programming such that $\delta = \sum_{i,j} A_{i,j} (\vec{v}_i^T \cdot \vec{v}_j)$. If $\rho \geq 0.5$, $\delta \geq \Delta \cdot M$ by choosing $\vec{v}_i = (1, 0, \cdots, 0)$ for every $i \notin S$

and $\vec{v}_i = (-\frac{1-\rho}{\rho}, 0, \cdots, 0)$ for every $i \in S$. But this is not a valid solution for the SDP when $\rho < 0.5$. However, $\delta \geq (\frac{\rho}{1-\rho})^2 \cdot \Delta \cdot M$ in this case by choosing choosing $\vec{v}_i = (\frac{\rho}{1-\rho}, 0, \cdots, 0)$ for every $i \notin S$ and $\vec{v}_i = (-1, 0, \cdots, 0)$ for every $i \in S$. Therefore $\delta \geq \min\{(\frac{\rho}{1-\rho})^2, 1\} \cdot \Delta M$. Without lose of generality, both δ and ΔM are $\geq \frac{1}{d}M$, otherwise a random subset is enough to achieve the desired approximation ratio.

The algorithm has two stages: first round $\vec{v}_i$ to $z_i \in [-1, 1]$ and keep $\sum_i z_i$ almost balanced, which is motivated by the work [AN04], then round z_i to x_i using the algorithm suggested by [CMM07].

Lemma 10.3.2. *There exists a polynomial time algorithm that given $\|\vec{v}_i\| \leq 1$ for every i, $\sum_i \vec{v}_i = \vec{0}$ and $\delta = \sum_j \|\frac{1}{d}\sum_{i\in\Gamma(j)} \vec{v}_i\|_2^2 \geq M/d$, it finds $z_i \in [-1, 1]$ for every i such that $|\sum_i z_i| = O(N/d)$ and $\sum_j (\frac{1}{d}\sum_{i\in\Gamma(j)} z_i)^2 \geq \Omega(\frac{\delta}{\log d})$.*

Proof. The algorithm works as follows:

1. Sample $\vec{g} \sim N(0, 1)^N$ and choose $t = 3\sqrt{\log d}$.
2. Let $\zeta_i = \langle g, \vec{v}_i \rangle$ for every $i = 1, 2, \cdots, n$.
3. If $\zeta_i > t$ or $\zeta_i < -t$, cut $\zeta_i = \pm t$ respectively.
4. $z_i = \zeta_i / t$.

It is convenient to analyze the approximation ratio in another set of vectors $\{\vec{u}_i | i \in [n]\}$ in a Hilbert space such that $\vec{u}_i(\vec{g}) = \langle \vec{v}_i, \vec{g} \rangle$ and $\langle \vec{u}_i, \vec{u}_j \rangle = E_{\vec{g}}[\langle \vec{u}_i, \vec{g} \rangle \cdot \langle \vec{g}, \vec{u}_j \rangle] = \langle \vec{v}_i, \vec{v}_j \rangle$. So $\sum_{i,j} A_{i,j}(\vec{u}_i^T \cdot \vec{u}_j) = \delta$ and $\sum_i \vec{u}_i = \vec{0}$ again. Let $\vec{u'}_i$ be the vector in the same Hilbert space by applying the cut operation with parameters t on $\vec{u}_i$. Namely $\vec{u'}_i(\vec{g}) = t$ (or $-t$) when $\vec{u}_i(\vec{g}) > t$ (or $< -t$), otherwise $\vec{u'}_i(\vec{g}) = \vec{u}_i(\vec{g}) \in [-t, t]$. Therefore the algorithm is as same as sampling a random point $\vec{g}$ and setting $z_i = \vec{u'}_i(\vec{g})/t$.

Fact 10.3.3. *For every i, $\|\vec{u'}_i - \vec{u}_i\|_1 = O(1/d^{4.5})$ and $\|\vec{u'}_i - \vec{u}_i\|_2^2 = O(1/d^4)$.*

The analysis uses the second fact to bound $\sum_{i,j} A_{i,j}((\vec{u}_i - \vec{u'}_i)^T \cdot \vec{u}_j) \leq O(m/d^2)$ as follows. Notice that A is a positive definite matrix and consider $\sum_{i,j} A_{i,j}(\vec{w}_i^T \cdot \vec{w'}_j)$ for any unit vectors $\vec{w}_i$ and $\vec{w'}_j$. It reaches the maximal value when $\vec{w}_i = \vec{w'}_i$ by the property of positive definite matrices. And $\sum_{i,j} A_{i,j}(\vec{w}_i^T \cdot \vec{w}_j) = \sum_{j\in[M]} \|\frac{1}{d}\sum_{i\in\Gamma(j)} \vec{w}_i\|_2^2$ is always bounded by M, because $\vec{w}_i$ are unit vectors. So $\sum_{i,j} A_{i,j}(\vec{w}_i^T \cdot \vec{w'}_j) \leq \max\{\|\vec{w}_1\|_2, \cdots, \|\vec{w}_n\|_2\} \cdot \max\{\|\vec{w'}_1\|_2, \cdots, \|\vec{w'}_n\|_2\} \cdot M$.

$$\begin{aligned}
&\sum_{i,j} A_{i,j}(\vec{u}_i^T \cdot \vec{u}_j) - \sum_{i,j} A_{i,j}(\vec{u'}_i^T \cdot \vec{u'}_j) \\
=&\sum_{i,j} A_{i,j}(\vec{u}_i^T \cdot (\vec{u}_j - \vec{u'}_j)) + \sum_{i,j} A_{i,j}((\vec{u}_i - \vec{u'}_i)^T \cdot \vec{u'}_j) \\
\leq& O(M/d^2)
\end{aligned}$$

Therefore $\sum_{i,j} A_{i,j}(\vec{u'}_i^T \cdot \vec{u'}_j) \geq 0.99 \sum_{i,j} A_{i,j}(\vec{u}_i^T \cdot \vec{u}_j) \geq 0.99M/d$. And it is upper bounded by $t^2 \cdot M$. So with probability at least $\frac{49}{dt^2}$, g satisfies $\sum_{i,j} A_{i,j} \cdot (\vec{u'}_i(g) \cdot \vec{u'}_j(g)) \geq .49\delta$. On the other hand, $|\sum_i \vec{u'}_i(g)| \geq N/d$ with probability at most $1/d^3$ from the first property $\|\sum_i \vec{u'}_i\|_1 \leq O(N/d^4)$.

Overall, with probability at least $\frac{.5}{dt^2} - 1/d^3$, z_i satisfies $|\sum z_i| = O(N/d)$ and $\sum_j (\frac{1}{d}\sum_{i\in\Gamma(j)} z_i)^2 = \Omega(\frac{\delta}{t^2}) = \Omega(\frac{\delta}{\log d})$. □

It is not difficult to verify that independently sampling $z_i \in \{-1, 1\}$ for every i according to its bias z_i will not reduce the objective value but keep the same bias overall i. Without lose of generality, let $z_i \in \{-1, 1\}$ from now on.

Lemma 10.3.4. *There exists a polynomial time algorithm that given z_i with $|\sum_i z_i| = O(N/d)$, outputs $x_i \in \{0, 1\}$ such that $\sum_i x_i = (1-\rho)(1 \pm 1/d^{1.5})N$ and $\sum_j 1_{\forall i\in\Gamma(j): x_i=1} \geq \Omega(d(1-\rho)^d) \cdot \sum_j (\frac{1}{d}\sum_{i\in\Gamma(j)} z_i)^2$.*

Proof. The algorithm works as follows:

1. $\delta = (1-\rho)\sqrt{2/d}$. Execute Step 2 or Step 3 with probability 0.5 and 0.5 separately.
2. For every $i \in [N]$, $x_i = 1$ with probability $1 - \rho + \delta z_i$.
3. For every $i \in [N]$, $x_i = 1$ with probability $1 - \rho - \delta z_i$.

Let $y_j = \frac{1}{d}\sum_{i\in\Gamma(j)} z_i$. The probability $x_i = 1$ for every i in $\Gamma(j)$ is

$$\begin{aligned}
&\frac{1}{2}\big((1-\rho+\delta)^{\frac{1+y_j}{2}d}\cdot(1-\rho-\delta)^{\frac{1-y_j}{2}d} + (1-\rho-\delta)^{\frac{1+y_j}{2}d}\cdot(1-\rho+\delta)^{\frac{1-y_j}{2}d}\big)\\
=&\frac{1}{2}(1-\rho+\delta)^{d/2}(1-\rho-\delta)^{d/2}\big((\frac{1-\rho+\delta}{1-\rho-\delta})^{y_j\cdot d/2} + (\frac{1-\rho-\delta}{1-\rho+\delta})^{y_j\cdot d/2}\big)\\
=&\frac{1}{2}(1-\rho)^d(1-\frac{\delta^2}{(1-\rho)^2})^{d/2}\cdot\cosh(y_j\cdot d/2\cdot\ln(\frac{1-\rho+\delta}{1-\rho-\delta}))\\
\geq&\frac{1}{2}(1-\rho)^d(1-2/d)^{d/2}\cdot 0.9\cdot\big(y_j\cdot d/2\cdot\ln(\frac{1-\rho+(1-\rho)\sqrt{2/d}}{1-\rho-(1-\rho)\sqrt{2/d}})\big)^2\\
\geq&\frac{1}{2}(1-\rho)^d(1-2/d)^{d/2}\cdot 0.9\cdot y_j^2\cdot(d/2)^2\cdot(\sqrt{2/d})^2\\
\geq&\Omega((1-\rho)^d\cdot y_j^2\cdot d).
\end{aligned}$$

At the same time, $\sum_i x_i$ is concentrated around $E[\sum_i x_i] = \sum_i(1-\rho\pm\delta z_i) = (1-\rho)N \pm \delta\sum_i z_i = (1-\rho)(1\pm 1/d^{1.5})N$ with very high probability. Therefore $\{x_1,\cdots,x_n\}$ satisfies $\sum_i x_i = (1-\rho)(1\pm 1/d^{1.5})N$ and $\sum_j 1_{\forall i\in\Gamma(j):x_i=1} \geq \Omega(d(1-\rho)^d)\cdot\sum_j y_j^2$ with constant probability. □

Proof of Theorem 10.3.1. Let δ be the value from SDP $(*)$, which is $\geq \min\{(\frac{\rho}{1-\rho})^2, 1\}\cdot\Delta$ from the analysis above. By Lemma 10.3.2, round v_i into $z_i \in [-1,1]$ such that $|\sum_i z_i| = O(N/d)$ and $\sum_j(\frac{1}{d}\sum_{i\in\Gamma(j)} z_i)^2 \geq \Omega(\frac{\delta}{\log d})$. By Lemma 10.3.4, round z_i into $x_i \in \{0,1\}$ such that $\sum_i x_i = (1-\rho)(1\pm 1/k^{1.5})N$ and $\sum_j 1_{\forall i\in\Gamma(j):x_i=1} \geq \Omega(d(1-\rho)^d\cdot\frac{\delta}{\log d})$.

Let $T = \{i | x_i = 0\}$. Then $|T| = \rho(1 \pm O(\frac{1}{k^{1.5}}))N$ and $\Gamma(T) \leq (1 - C \cdot \frac{\min\{(\frac{\rho}{1-\rho})^2, 1\}}{\log d} \cdot d(1-\rho)^d \cdot \Delta)M$ for some absolute constant C. At last, adjust the size of T by randomly adding or deleting $O(\frac{N}{k^{1.5}})$ vertices such that the size of T is ρN. Because at most $O(\frac{N}{k^{1.5}})$ vertices are added to T, with constant probability, $\Gamma(j) \cap T = \emptyset$ if $\Gamma(j) \cap T = \emptyset$ for a node $j \in [M]$ before the adjustment. Therefore $\Gamma(T) \leq (1 - C_0 \cdot \frac{\min\{(\frac{\rho}{1-\rho})^2, 1\}}{\log d} \cdot d(1-\rho)^d \cdot \Delta)M$ for some absolute constant C_0. $\square$

Appendices

A Omitted Proof in Chapter 3

We fix $z_1, \cdots, z_k$ to be complex numbers on the unit circle and use $Q(z)$ to denote the degree-k polynomial $\prod_{i=1}^{k}(z - z_i)$. We first bound the coefficients of $r_{n,k}(z)$ for $n \geq k$.

Lemma A.1. *Given $z_1, \cdots, z_k$, for any positive integer n, let $r_{n,k}(z) = \sum_{i=0}^{k-1} r_{n,k}^{(i)} \cdot z^i$ denote the residual polynomial of $r_{n,k} \equiv z^n \mod \prod_{j=1}^{k}(z - z_j)$. Then each coefficient in $r_{n,k}$ is bounded: $|r_{n,k}^{(i)}| \leq \binom{k-1}{i} \cdot \binom{n}{k-1}$ for every i.*

Proof. By definition, $r_{n,k}(z_i) = z_i^n$. From the polynomial interpolation, we have

$$r_{n,k}(z) = \sum_{i=1}^{k} \frac{\prod_{j \in [k] \setminus i}(z - z_j) z_i^n}{\prod_{j \in [k] \setminus i}(z_i - z_j)}.$$

Let $\mathsf{Sym}_{S,i}$ be the symmetry polynomial of $z_1, \cdots, z_k$ with degree i among subset $S \subseteq [k]$, i.e., $\mathsf{Sym}_{S,i} = \sum_{S' \subseteq \binom{S}{i}} \prod_{j \in S'} z_j$. Then the coefficients of z_l in $r_{n,k}(z)$ is

$$r_{n,k}(l) = (-1)^{k-1-l} \sum_{i=1}^{k} \frac{\mathsf{Sym}_{[k]\setminus i, k-1-l} \cdot z_i^n}{\prod_{j \in [k] \setminus i}(z_i - z_j)}.$$

We omit $(-1)^{k-1-l}$ in the rest of proof and use induction on n, k, and l to prove $|r_{n,k}^{(l)}| \leq \binom{k-1}{l}\binom{n}{k-1}$.

Base Case of n: For any $n < k$, from the definition, $r(z) = z^n$ and $|r_{n,k}^{(l)}| \leq 1$.

Suppose it is true for any $n < n_0$. We consider $r_{n_0,k}^{l}$ from now on. When $k = 1$, $r_{n,0} = z_1^n$ is bounded by 1 because z_1 is on the unit circle of $\mathbb{C}$.

Given n_0, suppose the induction hypothesis is true for any $k < k_0$ and any $l < k$. For $k = k_0$, we first prove that $|r_{n_0,k_0}^{(k_0-1)}| \leq \binom{n_0}{k_0-1}$ then prove that $|r_{n_0,k_0}^{(l)}| \leq \binom{k_0-1}{l}\binom{n_0}{k_0-1}$ for

$l = k_0 - 2, \cdots, 0.$

$$
\begin{aligned}
r_{n_0,k_0}^{(k_0-1)} &= \sum_{i=1}^{k_0} \frac{z_i^{n_0}}{\prod\limits_{j\in[k_0]\setminus i}(z_i - z_j)} \\
&= \sum_{i=1}^{k_0-1} \frac{z_i^{n_0}}{\prod\limits_{j\in[k_0]\setminus i}(z_i - z_j)} + \frac{z_{k_0}^{n_0}}{\prod\limits_{j\in[k_0]\setminus k_0}(z_{k_0} - z_j)} \\
&= \sum_{i=1}^{k_0-1} \frac{z_i^{n_0} - z_i^{n_0-1}z_{k_0} + z_i^{n_0-1}z_{k_0}}{\prod\limits_{j\in[k_0]\setminus i}(z_i - z_j)} + \frac{z_{k_0}^{n_0}}{\prod\limits_{j\in k_0\setminus k_0}(z_{k_0} - z_j)} \\
&= \sum_{i=1}^{k_0-1} \left(\frac{z_i^{n_0-1}}{\prod\limits_{j\in[k_0-1]\setminus i}(z_i - z_j)} + \frac{z_i^{n_0-1}z_{k_0}}{\prod\limits_{j\in k_0\setminus i}(z_i - z_j)} \right) + \frac{z_{k_0}^{n_0}}{\prod\limits_{j\in k_0\setminus k_0}(z_{k_0} - z_j)} \\
&= \left(\sum_{i=1}^{k_0-1} \frac{z_i^{n_0-1}}{\prod\limits_{j\in[k_0-1]\setminus i}(z_i - z_j)} \right) + \left(z_{k_0} \sum_{i=1}^{k_0} \frac{z_i^{n_0-1}}{\prod\limits_{j\in k_0\setminus i}(z_i - z_j)} \right) \\
&= r_{n_0-1,k_0-1}^{(k_0-2)} + z_{k_0} \cdot r_{n_0-1,k_0}^{(k_0-1)}
\end{aligned}
$$

Hence $|r^{(k_0-1)}_{n_0,k_0}| \le |r^{(k_0-2)}_{n_0-1,[k_0-1]}| + |r^{(k_0-1)}_{n_0-1,k_0}| \le \binom{n_0-2}{k_0-2} + \binom{n_0-2}{k_0-1} = \binom{n_0-1}{k_0-1}$. For $l < k_0 - 1$, we have

$$\begin{aligned}
r^{(l)}_{n_0,k_0} &= \sum_{i=1}^{k_0} \frac{\mathsf{Sym}_{[k_0]\setminus i,k_0-1-l} \cdot z_i^{n_0}}{\prod_{j\in[k_0]\setminus i}(z_i - z_j)} \quad \text{let } l' = k_0 - 1 - l \\
&= \sum_{i=1}^{k_0-1} \frac{\left(\mathsf{Sym}_{[k_0-1]\setminus i,l'} + \mathsf{Sym}_{[k_0-1]\setminus i,l'-1} \cdot z_{k_0}\right) z_i^{n_0}}{\prod_{j\in[k_0]\setminus i}(z_i - z_j)} + \frac{\mathsf{Sym}_{[k_0-1],l'} \cdot z_{k_0}^{n_0}}{\prod_{j<k_0}(z_{k_0} - z_j)} \\
&= \sum_{i=1}^{k_0-1} \frac{\mathsf{Sym}_{[k_0-1]\setminus i,l'} \cdot (z_i - z_{k_0}) z_i^{n_0-1} + \mathsf{Sym}_{[k_0-1]\setminus i,l'} \cdot z_{k_0} z_i^{n_0-1} + \mathsf{Sym}_{[k_0-1]\setminus i,l'-1} \cdot z_{k_0} z_i^{n_0}}{\prod_{j\in[k_0]\setminus i}(z_i - z_j)} \\
&\quad + \frac{\mathsf{Sym}_{[k_0-1],l'} \cdot z_{k_0}^{n_0}}{\prod_{j<k_0}(z_{k_0} - z_j)} \\
&= \sum_{i=1}^{k_0-1} \frac{\mathsf{Sym}_{[k_0-1]\setminus i,l'} \cdot (z_i - z_{k_0}) z_i^{n_0-1} + \mathsf{Sym}_{[k_0-1],l'} \cdot z_{k_0} z_i^{n_0-1}}{\prod_{j\in[k_0]\setminus i}(z_i - z_j)} + \frac{\mathsf{Sym}_{[k_0-1],l'} \cdot z_{k_0}^{n_0}}{\prod_{j<k_0}(z_{k_0} - z_j)} \\
&= \sum_{i=1}^{k_0-1} \frac{\mathsf{Sym}_{[k_0-1]\setminus i,l'} z_i^{n_0-1}}{\prod_{j\in[k_0-1]\setminus i}(z_i - z_j)} + \sum_{i=1}^{k_0-1} \frac{\mathsf{Sym}_{[k_0-1],l'} \cdot z_{k_0} z_i^{n_0-1}}{\prod_{j\in[k_0]\setminus i}(z_i - z_j)} + \frac{\mathsf{Sym}_{[k_0-1],l'} \cdot z_{k_0}^{n_0}}{\prod_{j<k_0}(z_{k_0} - z_j)} \\
&= r^{(l-1)}_{n_0-1,k_0-1} + \mathsf{Sym}_{[k_0-1],k_0-1-l} \cdot z_{k_0} \cdot r^{(k_0-1)}_{n_0-1,k_0}
\end{aligned}$$

By induction hypothesis, $|r^{(l)}_{n_0,k_0}| \le \binom{k_0-2}{l-1}\binom{n_0-1}{k_0-2} + \binom{k_0-1}{l}\binom{n_0-1}{k_0-1} \le \binom{k_0-1}{l}\binom{n_0}{k_0-1}$. □

Similarly, we could bound the coefficients of $z^{-n} \mod \prod_{j=1}^{k}(z - z_j)$.

Lemma A.2. *Given $z_1, \cdots, z_k$, for any integer n with $n \ge 0$, let $r_{-n,k}(z) = \sum_{i=0}^{k-1} r^{(i)}_{-n,k} \cdot z^i$ denote the residual polynomial of $r_{-n,k} \equiv z^{-n} \mod \prod_{j=1}^{k}(z - z_j)$. Then each coefficient in $r_{n,k}$ is bounded: $|r^{(i)}_{-n,k}| \le \binom{k-1}{i} \cdot \binom{n+k-1}{k-1}$ for every i.*

B Omitted Proof in Chapter 7

We finish the proof of the Leftover Hash Lemma 7.0.3 and the proof of Theorem 7.1.2 in this section.

Proof of Lemma 7.0.3. Given a subset Λ of size 2^k, we first consider

$$\begin{aligned}
&\sum_{h\in H}\sum_{\alpha\in[2^m]}(\Pr_{g\sim H,x\sim\Lambda}[h=g \text{ and } h(x)=\alpha]-\Pr_{g\sim H}[h=g]/2^m)^2\\
=&\sum_{h\in H}\sum_{\alpha\in[2^m]}\Pr_{g\sim H,x\sim\Lambda}[h=g \text{ and } h(x)=\alpha]^2\\
&-2\Pr_{g\sim H,x\sim\Lambda}[h=g \text{ and } h(x)=\alpha]\cdot\Pr_{g\sim H}[h=g]/2^m+(\Pr_{g\sim H}[h=g]/2^m)^2\\
=&\sum_{h\in H}\sum_{\alpha\in[2^m]}\Pr_{g\sim H,x\sim\Lambda}[h=g \text{ and } h(x)=\alpha]^2-\frac{1}{T\cdot 2^m}.
\end{aligned}$$

Notice that $\sum_{h,\alpha}\Pr_{g\sim H,x\sim\Lambda}[h=g \text{ and } h(x)=\alpha]^2$ is the collision probability

$$\begin{aligned}
&\Pr_{g\sim H,h\sim H,x\sim\Lambda,y\sim\Lambda}[h=g \text{ and } g(x)=h(y)]^2\\
=&\frac{1}{T}\cdot\Pr_{x\sim\Lambda,y\sim\Lambda}[x=y]+\frac{1}{T}\cdot\Pr_{x\sim\Lambda,y\sim\Lambda}[x\neq y]\cdot\Pr_{h\sim H}[h(x)=h(y)|x\neq y]\\
=&\frac{1}{T}\big(2^{-k}+(2^{-m}+2^{-k})(1-2^{-k})\big)\\
\leq&\frac{1}{T}\big(2\cdot 2^{-k}+2^{-m}\big).
\end{aligned}$$

Thus $\sum_{h\in H}\sum_{\alpha\in[2^m]}(\Pr_{g\sim H,x\sim\Lambda}[h=g \text{ and } h(x)=\alpha]-\Pr_{g\sim H}[h=g]/2^m)^2\leq\frac{2}{T\cdot 2^k}$. From the Cauchy-Schwartz inequality,

$$\sum_{h\in H}\sum_{\alpha\in[2^m]}|\Pr_{g\sim H,x\sim\Lambda}[h=g \text{ and } h(x)=\alpha]-\Pr_{g\sim H}[h=g]/2^m|\leq\sqrt{T\cdot 2^m}\cdot\sqrt{\frac{2}{T\cdot 2^k}}\leq\sqrt{2}\cdot 2^{-\frac{k-m}{2}}\leq\sqrt{2}\epsilon.$$

Thus $\mathsf{Ext}(x,a)=h_a(x)$ is a strong extractor error $\frac{\sqrt{2}\epsilon}{2}$ (in statistical distance). □

Then we apply symmetrization and Gaussianization to prove Theorem 7.1.2.

Proof of Theorem 7.1.2. We first symmetrize it by

$$\begin{aligned}
&\mathop{\mathbb{E}}_{x_1,\cdots,x_n}\left[\max_{\Lambda}\sum_{j=1}^{n} f(\Lambda,x_j)\right]\\
=&\mathop{\mathbb{E}}_{x_1,\cdots,x_n}\left[\max_{\Lambda}\left(\sum_{j=1}^{n} f(\Lambda,x_j)-\mathop{\mathbb{E}}_{x'_1,\cdots,x'_n}[\sum_{j=1}^{n} f(\Lambda,x'_j)]+\mathop{\mathbb{E}}_{x'_1,\cdots,x'_n}[\sum_{j=1}^{n} f(\Lambda,x'_j)]\right)\right]\\
\le&\max_{\Lambda}\mathop{\mathbb{E}}_{x'}\left[\sum_{j=1}^{n} f(\Lambda,x'_j)\right]+\mathop{\mathbb{E}}_{x}\left[\max_{\Lambda}\left|\sum_{j=1}^{n} f(\Lambda,x_j)-\mathop{\mathbb{E}}_{x'}[\sum_{j=1}^{n} f(\Lambda,x'_j)]\right|\right].
\end{aligned}$$

Then we apply Gaussianization on the second term.

Claim B.1. *Let $g=(g_1,\cdots,g_n)$ denote the Gaussian vector sampled from $N(0,1)^n$,*

$$\mathop{\mathbb{E}}_{x}\left[\max_{\Lambda}\left|\sum_{j=1}^{n} f(\Lambda,x_j)-\mathop{\mathbb{E}}_{x'}[\sum_{j=1}^{n} f(\Lambda,x'_j)]\right|\right]\le\sqrt{2\pi}\cdot\mathop{\mathbb{E}}_{x}\left[\mathop{\mathbb{E}}_{g}\left[\max_{\Lambda}\left|\sum_{j=1}^{n} f(\Lambda,x_j)g_j\right|\right]\right].$$

Proof. We first use the convexity of the $|\cdot|$ function to move $\mathop{E}_{x'}$ to the left hand side:

$$\begin{aligned}
\mathop{\mathbb{E}}_{x}\left[\max_{\Lambda}\left|\sum_{j=1}^{n} f(\Lambda,x_j)-\mathop{\mathbb{E}}_{x'}[\sum_{j=1}^{n} f(\Lambda,x'_j)]\right|\right]&\le\mathop{\mathbb{E}}_{x}\left[\max_{\Lambda}\mathop{\mathbb{E}}_{x'}\left|\sum_{j=1}^{n} f(\Lambda,x_j)-\sum_{j=1}^{n} f(\Lambda,x'_j)\right|\right]\\
&\left(\text{use }\max_i\mathop{\mathbb{E}}_{G}[G_i]\le\mathop{\mathbb{E}}_{G}[\max_i G_i]\text{ to move }\mathop{\mathbb{E}}_{x'}\text{ out}\right)\\
&\le\mathop{\mathbb{E}}_{x,x'}\left[\max_{\Lambda}\left|\sum_{j=1}^{n} f(\Lambda,x_j)-\sum_{j=1}^{n} f(\Lambda,x'_j)\right|\right]
\end{aligned}$$

Then we Gaussianize it using the fact $\mathbb{E}[|g_j|]=\sqrt{2/\pi}$. Let g denote a sequence of n

independent Gaussian random variables. We have

$$\begin{aligned}
\mathop{\mathbb{E}}_{x,x'}\left[\max_{\Lambda}\left|\sum_{j=1}^{n} f(\Lambda,x_j)-\sum_{j=1}^{n} f(\Lambda,x'_j)\right|\right] &= \sqrt{\pi/2}\mathop{\mathbb{E}}_{x,x'}\left[\max_{\Lambda}\left|\sum_{j=1}^{n}\big(f(\Lambda,x_j)-f(\Lambda,x'_j)\big)\cdot\mathop{\mathbb{E}}_{g_j}|g_j|\right|\right] \\
&\left(\text{use the convexity of } |\cdot| \text{ to move } \mathop{\mathbb{E}}_{g}\right) \\
&\le \sqrt{\pi/2}\mathop{\mathbb{E}}_{x,x'}\left[\max_{\Lambda}\mathop{\mathbb{E}}_{g}\left|\sum_{j=1}^{n}\big(f(\Lambda,x_j)-f(\Lambda,x'_j)\big)\cdot|g_j|\right|\right] \\
&\left(\text{use } \max_i \mathop{\mathbb{E}}_{G}[G_i]\le \mathop{\mathbb{E}}_{G}[\max_i G_i] \text{ to move } \mathop{\mathbb{E}}_{g} \text{ out}\right) \\
&\le \sqrt{\pi/2}\mathop{\mathbb{E}}_{x,x'}\mathop{\mathbb{E}}_{g}\left[\max_{\Lambda}\left|\sum_{j=1}^{n}\big(f(\Lambda,x_j)-f(\Lambda,x'_j)\big)\cdot|g_j|\right|\right] \\
&\left(\text{use the symmetry of } f(\Lambda,x_j)-f(\Lambda,x'_j)\right) \\
&= \sqrt{\pi/2}\mathop{\mathbb{E}}_{g}\mathop{\mathbb{E}}_{x,x'}\left[\max_{\Lambda}\left|\sum_{j=1}^{n}\big(f(\Lambda,x_j)-f(\Lambda,x'_j)\big)\cdot g_j\right|\right] \\
&\left(\text{use the triangle inequality}\right) \\
&\le \sqrt{\pi/2}\mathop{\mathbb{E}}_{x,x'}\mathop{\mathbb{E}}_{g}\left[\max_{\Lambda}\left|\sum_{j=1}^{n} f(\Lambda,x_j)g_j\right|+\max_{\Lambda}\left|-\sum_{j=1}^{n} f(\Lambda,x'_j)g_j\right|\right] \\
&\left(\text{use the symmetry of } g_j\right) \\
&\le \sqrt{2\pi}\mathop{\mathbb{E}}_{x}\mathop{\mathbb{E}}_{g}\left[\max_{\Lambda}\left|\sum_{j=1}^{n} f(\Lambda,x_j)g_j\right|\right].
\end{aligned}$$

□

Remark B.2. *We use the independence between $x_1,\cdots,x_n$ in the third last step.*

□

C Omitted Proof in Chapter 8

We apply an induction from $i = 0$ to $i = k$. The base case $i = 0$ is true because there are at most m balls.

Suppose it is true for $i = l < d$. For a fixed bin $j \in [n^{1-\frac{1}{2^l}}]$, there are at most $(1+\beta)^l n^{\frac{1}{2^l}} \cdot \frac{m}{n}$ balls. With out loss of generality, we assume there are exactly $s = (1+\beta)^l n^{\frac{1}{2^l}} \cdot \frac{m}{n}$ balls from the induction hypothesis. Under the hash function h_{l+1}, we allocate these balls into $t = n^{\frac{1}{2^{l+1}}}$ bins.

We fix one bin in h_{l+1} and prove that this bin receives at most

$$(1+\beta)s/t = (1+\beta) \cdot (1+\beta)^l n^{\frac{1}{2^l}} / n^{\frac{1}{2^{l+1}}} \cdot \frac{m}{n} = (1+\beta)^{l+1} n^{\frac{1}{2^{l+1}}} \cdot \frac{m}{n}$$

balls with probability $\leq 2n^{-c-2}$ in a $\log^3 n$-wise δ_1-biased space for $\delta_1 = 1/\text{poly}(n)$ with a sufficiently large polynomial. We use $X_i \in \{0, 1\}$ to denote the ith ball is in the bin or not. Hence $\mathbb{E}[\sum_{i\in[s]} X_i] = s/t$.

For convenience, we use $Y_i = X_i - \mathbb{E}[X_i]$. Hence $Y_i = 1 - 1/t$ w.p. $1/t$, o.w. $Y_i = -1/t$. Notice that $\mathbb{E}[Y_i] = 0$ and $|\mathbb{E}[Y_i^l]| \leq 1/t$ for any $l \geq 2$.

We choose $b = 2^l \cdot \beta = O(\log n)$ for a large even number β and compute the bth moment of $\sum_{i\in[s]} Y_i$ as follows.

$$\begin{aligned}
\Pr[|\sum_{i\in[s]} X_i| > (1+\beta)s/t] &\leq \mathop{\mathbb{E}}_{\delta_1\text{-}biased}[(\sum_i Y_i)^b]/(\beta s/t)^b \\
&\leq \frac{\mathbb{E}_U[(\sum_i Y_i)^b] + \delta_1 \cdot s^{2b}t^b}{(\beta s/t)^b} \\
&\leq \frac{\sum_{i_1,\ldots,i_b} \mathbb{E}[Y_{i_1} \cdot Y_{i_2} \cdot \cdots \cdot Y_{i_b}] + \delta_1 \cdot s^{3b}}{(\beta s/t)^b} \\
&\leq \frac{\sum_{j=1}^{b/2} \binom{b-j-1}{j-1} \cdot \frac{b!}{2^{b/2}} \cdot s^j (1/t)^j + \delta_1 \cdot s^{3b}}{(\beta s/t)^b} \\
&\leq \frac{2^{b/2} b! \cdot (s/t)^{b/2} + \delta_1 \cdot s^{3b}}{(\beta s/t)^b}
\end{aligned}$$

Because $s/t \geq n^{\frac{1}{2^k}} \geq \log^3 n, b \leq \beta 2^k \leq \frac{\beta \log n}{3 \log\log n} \leq (s/t)^{1/3}$ and $\beta = (\log n)^{-0.2} < (s/t)^{0.1}$, we simplify it to

$$\left(\frac{2b^2 \cdot s/t}{(\beta s/t)^2}\right)^{b/2} + \delta_1 \cdot s^{3b} \leq \left(\frac{2(s/t)^{2/3} \cdot s/t}{(s/t)^{1.8}}\right)^{b/2} + \delta_1 \cdot s^{3b}$$
$$=(s/t)^{(-0.1)\cdot b/2} + \delta_1 \cdot s^{3b} \leq (n^{\frac{2}{2^{l+1}}})^{-0.1\cdot\beta 2^l} + \delta_1 (n^{\frac{3}{2^l}})^{3\beta\cdot 2^l} = n^{-c-2} + \delta_1 n^{9\beta} \leq 2n^{-c-2}.$$

Finally we choose $\beta = 40(c+2) = O(1)$ and $\delta_1 = n^{-9\beta-c-2}$ to finish the proof.

Bibliography

[ABG13] Per Austrin, Siavosh Benabbas, and Konstantinos Georgiou. Better Balance by Being Biased: A 0.8776-Approximation for Max Bisection. In *Proceedings of the 24th Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '13, pages 277–294, 2013. 13

[ABKU99] Yossi Azar, Andrei Z. Broder, Anna R. Karlin, and Eli Upfal. Balanced allocations. *SIAM J. Comput.*, 29(1):180–200, September 1999. 11, 12, 105

[AGHP90] N. Alon, O. Goldreich, J. Hastad, and R. Peralta. Simple construction of almost k-wise independent random variables. In *Proceedings of the 31st Annual Symposium on Foundations of Computer Science*. IEEE Computer Society, 1990. 107

[AGK+11] Noga Alon, Gregory Gutin, EunJung Kim, Stefan Szeider, and Anders Yeo. Solving MAX-r-SAT Above a Tight Lower Bound. *Algorithmica*, 61(3):638–655, 2011. 132, 134

[AN04] Noga Alon and Assaf Naor. Approximating the cut-norm via grothendieck's inequality. In *Proceedings of the 36th Annual ACM Symposium on Theory of Computing, Chicago, IL, USA, June 13-16, 2004*, pages 72–80, 2004. 197

[Bar14] Boaz Barak. Sums of squares upper bounds, lower bounds, and open questions. http://www.boazbarak.org/sos/, 2014. Page 39. Accessed: October 28, 2014. 180

[BBH+14] Boaz Barak, Fernando G.S.L. Brandao, Aram W. Harrow, Jonathan Kelner, David Steurer, and Yuan Zhou. Hypercontractivity, Sum-of-squares Proofs, and Their Applications. In *Proceedings of the 44th Annual ACM Symposium on Theory of Computing*, STOC '14, pages 307–326, 2014. 132

[BCG+14] Petros Boufounos, Volkan Cevher, Anna C Gilbert, Yi Li, and Martin J Strauss. What's the frequency, Kenneth?: Sublinear Fourier sampling off the grid. In *Algorithmica(A preliminary version of this paper appeared in the Proceedings of RANDOM/APPROX 2012, LNCS 7408, pp. 61-72)*, pages 1–28. Springer, 2014. 3

[BCV+12] Aditya Bhaskara, Moses Charikar, Aravindan Vijayaraghavan, Venkatesan Guruswami, and Yuan Zhou. Polynomial integrality gaps for strong sdp relaxations of densest k-subgraph. In *Proceedings of the 23rd Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 388–405. SIAM, 2012. 191

[BDMI13] Christos Boutsidis, Petros Drineas, and Malik Magdon-Ismail. Near-optimal coresets for least-squares regression. *IEEE transactions on information theory*, 59(10):6880–6892, 2013. 6

[BDT17] Avraham Ben-Aroya, Dean Doron, and Amnon Ta-Shma. An efficient reduction from two-source to non-malleable extractors: achieving near-logarithmic min-entropy. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing, STOC 2017*, pages 1185–1194, 2017. 8

[Ber97] Bonnie Berger. The Fourth Moment Method. *SIAM J. Comput.*, 26(4):1188–1207, August 1997. 134

[BGGP12] Itai Benjamini, Ori Gurel-Gurevich, and Ron Peled. On k-wise independent

distributions and boolean functions. http://arxiv.org/abs/1207.0016, 2012. Accessed: October 28, 2014. 182

[BGH+12] Boaz Barak, Parikshit Gopalan, Johan Hastad, Raghu Meka, Prasad Raghavendra, and David Steurer. Making the long code shorter. In *Proceedings of the 2012 IEEE 53rd Annual Symposium on Foundations of Computer Science*, pages 370–379, Washington, DC, USA, 2012. IEEE Computer Society. 178

[BKS+10] B. Barak, G. Kindler, R. Shaltiel, B. Sudakov, and A. Wigderson. Simulating independence: New constructions of condensers, ramsey graphs, dispersers, and extractors. *J. ACM*, 57(4):20:1–20:52, May 2010. 177

[BM86] Y. Bresler and A. Macovski. Exact maximum likelihood parameter estimation of superimposed exponential signals in noise. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 34(5):1081–1089, Oct 1986. 3

[BMRV00] H. Buhrman, P. B. Miltersen, J. Radhakrishnan, and S. Venkatesh. Are bitvectors optimal? In *Proceedings of the 32nd Annual ACM Symposium on Theory of Computing*, STOC, pages 449–458, New York, NY, USA, 2000. ACM. 177

[Bon70] Aline Bonami. Étude des coefficients Fourier des fonctions de $L^p(G)$. *Annales de l'Institut Fourier*, 20(2):335–402, 1970. 135

[BOT02] Andrej Bogdanov, Kenji Obata, and Luca Trevisan. A lower bound for testing 3-colorability in bounded-degree graphs. In *Proceedings of the 43rd Symposium on Foundations of Computer Science*, pages 93–102, 2002. 177

[BSS12] Joshua Batson, Daniel A Spielman, and Nikhil Srivastava. Twice-ramanujan sparsifiers. *SIAM Journal on Computing*, 41(6):1704–1721, 2012. 6, 54, 67

[Cha00] Bernard Chazelle. *The Discrepancy Method*. Cambridge University Press, 2000. 83

[Cha13] Siu On Chan. Approximation resistance from pairwise independent subgroups. In *Proceedings of the Forty-fifth Annual ACM Symposium on Theory of Computing*, STOC '13, pages 447–456, New York, NY, USA, 2013. ACM. 187, 188

[Che52] Herman Chernoff. A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *The Annals of Mathematical Statistics*, 23:493–507, 1952. 17

[Che16] Xue Chen. Integrality gaps and approximation algorithms for dispersers and bipartite expanders. In *SODA*, 2016. 14

[Che17] Xue Chen. Derandomized allocation process. Manuscript (https://arxiv.org/abs/1702.03375), 2017. 14

[CJM15] Robert Crowston, Mark Jones, and Matthias Mnich. Max-cut parameterized above the edwards-erdős bound. *Algorithmica*, 72(3):734–757, 2015. 133

[CKNS15] Kamalika Chaudhuri, Sham M Kakade, Praneeth Netrapalli, and Sujay Sanghavi. Convergence rates of active learning for maximum likelihood estimation. In *Advances in Neural Information Processing Systems*, pages 1090–1098, 2015. 7

[CKPS16] Xue Chen, Daniel M. Kane, Eric Price, and Zhao Song. Fourier-sparse interpolation without a frequency gap. In *Foundations of Computer Science(FOCS), 2016 IEEE 57th Annual Symposium on*, 2016. 4, 5, 14, 41

[CL16] Eshan Chattopadhyay and Xin Li. Extractors for sumset sources. In *Proceedings of the 48th Annual ACM SIGACT Symposium on Theory of Computing, STOC 2016, Cambridge, MA, USA, June 18-21, 2016*, pages 299–311, 2016. 10

[CMM07] Moses Charikar, Konstantin Makarychev, and Yury Makarychev. Near-optimal algorithms for maximum constraint satisfaction problems. In *Proceedings of the 18th Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '07, pages 62–68, 2007. 197

[CP17] Xue Chen and Eric Price. Condition number-free query and active learning of linear families. Manuscript (`https://arxiv.org/abs/1711.10051`), 2017. 4, 5, 6, 14

[CRSW13] L. Elisa Celis, Omer Reingold, Gil Segev, and Udi Wieder. Balls and bins: Smaller hash families and faster evaluation. *SIAM J. Comput.*, 42(3):1030–1050, 2013. 11, 12, 105, 114, 115, 116, 127

[CRVW02] Michael Capalbo, Omer Reingold, Salil Vadhan, and Avi Wigderson. Randomness conductors and constant-degree lossless expanders. In *Proceedings of the 34th Annual ACM STOC*, pages 659–668. ACM, 2002. 177

[CW79] J. Lawrence Carter and Mark N. Wegman. Universal classes of hash functions (extended abstract). *JOURNAL OF COMPUTER AND SYSTEM SCIENCES*, 18:143–154, 1979. 80, 81

[CZ17] Xue Chen and Yuan Zhou. Parameterized algorithms for constraint satisfaction problems above average with global cardinality constraints. In *SODA*, 2017. 14

[CZ18] Xue Chen and David Zuckerman. Existence of simple extractors. *manuscript*, 2018. 8, 14

[DHKP97] Martin Dietzfelbinger, Torben Hagerup, Jyrki Katajainen, and Martti Penttonen. A reliable randomized algorithm for the closest-pair problem. *J. Algorithms*, 25(1):19–51, October 1997. 81

[DMM08] Petros Drineas, Michael W Mahoney, and S Muthukrishnan. Relative-error cur matrix decompositions. *SIAM Journal on Matrix Analysis and Applications*, 30(2):844–881, 2008. 6

[Dun10] Mark Dunster. *Legendre and Related Functions*. Handbook of Mathematical Functions, Cambridge University Press, 2010. 30

[Fan61] Robert Fano. *Transmission of information; a statistical theory of communications*. Cambridge, Massachusetts, M.I.T. Press, 1961. 75

[FJ95] Alan Frieze and Mark Jerrum. Improved approximation algorithms for max k-cut and max bisection. In *Proceedings of the 4th International Conference on Integer Programming and Combinatorial Optimization*, pages 1–13, 1995. 13

[FL06] Uriel Feige and Michael Langberg. The RPR2 Rounding Technique for Semidefinite Programs. *J. Algorithms*, 60(1):1–23, July 2006. 13

[FL12] Albert Fannjiang and Wenjing Liao. Coherence pattern-guided compressive sensing with unresolved grids. *SIAM Journal on Imaging Sciences*, 5(1):179–202, 2012. 3

[FM16] Yuval Filmus and Elchanan Mossel. Harmonicity and Invariance on Slices of the Boolean Cube. In *Computational Complexity Conference 2016*, CCC 2016, 2016. 133

[GKM15] Parikshit Gopalan, Daniek Kane, and Raghu Meka. Pseudorandomness via the discrete fourier transform. In *IEEE 56th Annual Symposium on Foundations of Computer Science, FOCS 2015*, pages 903–922, 2015. 106

[GMR+11] Venkatesan Guruswami, Yury Makarychev, Prasad Raghavendra, David Steurer, and Yuan Zhou. Finding Almost-Perfect Graph Bisections. In *Proceedings of the 2nd Symposium on Innovations in Computer Science*, ICS '11, pages 321–337, 2011. 13

[God10] Chris Godsil. *Association Schemes*. http://www.math.uwaterloo.ca/~cgodsil/pdfs/assoc2.pdf, 2010. Last accessed: November 6, 2015. 138

[Gri01a] Dima Grigoriev. Linear lower bound on degrees of positivstellensatz calculus proofs for the parity. *Theoretical Computer Science*, 259:613 – 622, 2001. 177, 187

[Gri01b] Dmitry Grigoriev. Complexity of Positivstellensatz proofs for the knapsack. *Computational Complexity*, 10(2):139–154, 2001. 140

[GSZ14] Venkatesan Guruswami, Ali Kemal Sinop, and Yuan Zhou. Constant factor lasserre integrality gaps for graph partitioning problems. *SIAM Journal on Optimization*, 24(4):1698 – 1717, 2014. 181

[GUV09a] Venkatesan Guruswami, Christopher Umans, and Salil Vadhan. Unbalanced expanders and randomness extractors from parvaresh–vardy codes. *J. ACM*, 56(4):20:1–20:34, July 2009. 10

[GUV09b] Venkatesan Guruswami, Christopher Umans, and Salil Vadhan. Unbalanced expanders and randomness extractors from parvaresh–vardy codes. *J. ACM*, 56(4):20:1–20:34, July 2009. 177

[GY10] Gregory Gutin and Anders Yeo. Note on maximal bisection above tight lower bound. *Inf. Process. Lett.*, 110(21):966–969, 2010. 132, 133

[Har28] Ralph Hartley. Transmission of information. *Bell System Technical Journal*, 1928. 75

[Haz01] Michiel Hazewinkel. *Gram matrix*. Encyclopedia of Mathematics, Springer, 2001. 36

[HIKP12] Haitham Hassanieh, Piotr Indyk, Dina Katabi, and Eric Price. Nearly optimal sparse Fourier transform. In *Proceedings of the forty-fourth annual ACM symposium on Theory of computing*. ACM, 2012. 2

[HZ02] Eran Halperin and Uri Zwick. A Unified Framework for Obtaining Improved Approximation Algorithms for Maximum Graph Bisection Problems. *Random Struct. Algorithms*, 20(3):382–402, May 2002. 13

[IK14] Piotr Indyk and Michael Kapralov. Sample-optimal Fourier sampling in any constant dimension. In *Foundations of Computer Science (FOCS), 2014 IEEE 55th Annual Symposium on*, pages 514–523. IEEE, 2014. 2

[ILL89] R. Impagliazzo, L. A. Levin, and M. Luby. Pseudo-random generation from one-way functions. In *Proceedings of the Twenty-first Annual ACM Symposium on Theory of Computing*, STOC '89, pages 12–24, New York, NY, USA, 1989. ACM. 10, 79, 80, 81

[Kah95] Nabil Kahale. Eigenvalues and expansion of regular graphs. *J. ACM*, 42(5):1091–1106, September 1995. 177

[Kho02] Subhash Khot. On the Power of Unique 2-prover 1-round Games. In *Proceedings of the 34th Annual ACM Symposium on Theory of Computing*, STOC '02, pages 767–775, New York, NY, USA, 2002. ACM. 13

[KKL88] J. Kahn, G. Kalai, and N. Linial. The Influence of Variables on Boolean Functions. In *Proceedings of the 29th Annual Symposium on Foundations of Computer Science*, FOCS '88, pages 68–80, 1988. 132

[KV05] Subhash Khot and Nisheeth K. Vishnoi. The unique games conjecture, integrality gap for cut problems and embeddability of negative type metrics into l_1. In *46th Annual IEEE Symposium on Foundations of Computer Science, 23-25 October 2005, Pittsburgh, PA, USA, Proceedings*, pages 53–62, 2005. 178

[Las02] Jean B. Lasserre. An explicit equivalent positive semidefinite program for nonlinear 0-1 programs. *SIAM J. on Optimization*, 12(3):756–769, March 2002. 180, 181

[Li16] Xin Li. Improved two-source extractors, and affine extractors for polylogarithmic entropy. In *IEEE 57th Annual Symposium on Foundations of Computer Science, FOCS 2016, 9-11 October 2016, Hyatt Regency, New Brunswick, New Jersey, USA*, pages 168–177, 2016. 10

[LS15] Yin Tat Lee and He Sun. Constructing linear-sized spectral sparsification in almost-linear time. In *Proceedings of the 2015 IEEE 56th Annual Symposium on Foundations of Computer Science (FOCS)*, FOCS '15, pages 250–269. IEEE Computer Society, 2015. 6, 54, 55, 56, 67

[LT91] M. Ledoux and M. Talagrand. *Probability in Banach spaces*. Springer, 1991. 82, 91

[LY98] Tzong-Yau Lee and Horng-Tzer Yau. Logarithmic Sobolev inequality for some models of random walks. *Ann. Prob.*, 26(4):1855–1873, 1998. 133

[Mah11] Michael W Mahoney. Randomized algorithms for matrices and data. *Foundations and Trends® in Machine Learning*, 3(2):123–224, 2011. 6

[MI10] Malik Magdon-Ismail. Row sampling for matrix algorithms via a non-commutative bernstein bound. *arXiv preprint arXiv:1008.0587*, 2010. 6

[MMZ15] Konstantin Makarychev, Yury Makarychev, and Yuan Zhou. Satisfiability of Ordering CSPs Above Average Is Fixed-Parameter Tractable. In *Proceedings of the 56th Annual IEEE Symposium on Foundations of Computer Science*, FOCS '15, pages 975–993, 2015. 132

[Moi15] Ankur Moitra. The threshold for super-resolution via extremal functions. In *STOC*, 2015. 3, 4

[MOO05] Elchanan Mossel, Ryan O'Donnell, and Krzysztof Oleszkiewicz. Noise stability of functions with low influences: invariance and optimality. In *Proceedings of the 46th Annual IEEE Symposium on Foundations of Computer Science*, FOCS '05, pages 21–30, 2005. 132

[MOO10] Elchanan Mossel, Ryan O'Donnell, and Krzysztof Oleszkiewicz. Noise stability of functions with low influences: invariance and optimality. *Annals of Mathematics*, 171(1):295–341, 2010. 132

[MRRR14] Raghu Meka, Omer Reingold, Guy N. Rothblum, and Ron D. Rothblum. Fast pseudorandomness for independence and load balancing - (extended abstract). In *Automata, Languages, and Programming - 41st International Colloquium, ICALP 2014, Proceedings*, pages 859–870, 2014. 105, 116

[MZ12] Matthias Mnich and Rico Zenklusen. Bisections above tight lower bounds. In *the 38th International Workshop on Graph-Theoretic Concepts in Computer Science*, WG '12, pages 184–193, 2012. 132, 133

[NN90] J. Naor and M. Naor. Small-bias probability spaces: Efficient constructions and applications. In *Proceedings of the Twenty-second Annual ACM Symposium on Theory of Computing*, STOC '90, pages 213–223. ACM, 1990. 107

[NZ96] Noam Nisan and David Zuckerman. Randomness is linear in space. *J. Comput. Syst. Sci.*, 52(1):43–52, 1996. 7

[O'D14] Ryan O'Donnell. *Analysis of Boolean Functions*. Cambridge University Press, 2014. 132, 134, 135

[Owe13] Art B. Owen. *Monte Carlo theory, methods and examples*. 2013. 6

[PS15] Eric Price and Zhao Song. A robust sparse Fourier transform in the continuous setting. In *Foundations of Computer Science (FOCS), 2015 IEEE 56th Annual Symposium on*, pages 583–600. IEEE, 2015. 3

[Rot13] Thomas Rothvoß. The lasserre hierarchy in approximation algorithms. http://www.math.wash 2013. Accessed: October 28, 2014. 180

[RPK86] Robert Roy, Arogyaswami Paulraj, and Thomas Kailath. Esprit–a subspace rotation approach to estimation of parameters of cisoids in noise. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 34(5):1340–1342, 1986. 3

[RRV99] Ran Raz, Omer Reingold, and Salil Vadhan. Error reduction for extractors. In *Proceedings of the 40th Annual Symposium on the Foundations of Computer Science*, New York, NY, October 1999. IEEE. 11

[RRW14] Omer Reingold, Ron D. Rothblum, and Udi Wieder. Pseudorandom graphs in data structures. In *Automata, Languages, and Programming - 41st International Colloquium, ICALP 2014, Proceedings*, pages 943–954, 2014. 12

[RS10] Prasad Raghavendra and David Steurer. Graph expansion and the unique games conjecture. In *Proceedings of the 42nd ACM STOC*, pages 755–764, 2010. 13

[RST10] Prasad Raghavendra, David Steurer, and Prasad Tetali. Approximations for the isoperimetric and spectral profile of graphs and related parameters. In *Proceedings of the 42nd ACM STOC*, pages 631–640, 2010. 13

[RST12] Prasad Raghavendra, David Steurer, and Madhur Tulsiani. Reductions between expansion problems. In *Proceedings of the 27th Conference on Computational Complexity*, pages 64–73, 2012. 13

[RT00] Jaikumar Radhakrishnan and Amnon Ta-Shma. Bounds for dispersers, extractors, and depth-two superconcentrators. *SIAM J. Discrete Math.*, 13(1):2–24, 2000. 8

[RT12] Prasad Raghavendra and Ning Tan. Approximating CSPs with global cardinality constraints using SDP hierarchies. In *Proceedings of the 23rd Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '12, pages 373–387, 2012. 13

[RV08] Mark Rudelson and Roman Vershynin. On sparse reconstruction from fourier and gaussian measurements. *Communications on Pure and Applied Mathematics*, 61(8):1025–1045, 2008. 91, 92

[RW14] Atri Rudra and Mary Wootters. Every list-decodable code for high noise has abundant near-optimal rate puncturings. In *STOC*, 2014. 82

[Sch81] Ralph Otto Schmidt. A signal subspace approach to multiple emitter location spectral estimation. *Ph. D. Thesis, Stanford University*, 1981. 3

[Sch08] Grant Schoenebeck. Linear level lasserre lower bounds for certain k-csps. In *Proceedings of the 2008 49th Annual IEEE Symposium on Foundations of Computer Science*, FOCS '08, pages 593–602, Washington, DC, USA, 2008. 187

[Sha49] Claude Shannon. Communication in the presence of noise. *Proc. Institute of Radio Engineers*, 37(1):10–21, 1949. 75

[Sha02] Ronen Shaltiel. Recent developments in explicit constructions of extractors. *Bulletin of the EATCS*, 77:67–95, 2002. 178

[Sip88] Michael Sipser. Expanders, randomness, or time versus space. *J. Comput. Syst. Sci.*, 36(3):379–383, June 1988. 177

[SM14] Sivan Sabato and Remi Munos. Active regression by stratification. In *Advances in Neural Information Processing Systems*, pages 469–477, 2014. 7

[SS96] Michael Sipser and Daniel Spielman. Expander codes. 6:1710–1722, 1996. 177

[SS11] Daniel A Spielman and Nikhil Srivastava. Graph sparsification by effective resistances. *SIAM Journal on Computing*, 40(6):1913–1926, 2011. 6

[SSS95] Jeanette P. Schmidt, Alan Siegel, and Aravind Srinivasan. Chernoff-hoeffding bounds for applications with limited independence. *SIAM J. Discret. Math.*, 8(2):223–250, May 1995. 108, 129

[Sti94] D. R. Stinson. Combinatorial techniques for universal hashing. *Journal of Computer and System Sciences*, 48(2):337–346, 1994. 116

[SU05] Ronen Shaltiel and Christopher Umans. Simple extractors for all min-entropies and a new pseudorandom generator. *J. ACM*, 52(2):172–216, March 2005. 10

[SV86] Miklos Santha and Umesh V. Vazirani. Generating quasi-random sequences from slightly-random sources. *J. Comput. System Sci.*, (33):75–87, 1986. 7

[SWZ17] Zhao Song, David P Woodruff, and Peilin Zhong. Relative error tensor low rank approximation. *arXiv preprint arXiv:1704.08246*, 2017. 6

[Ta-02] Amnon Ta-Shma. Almost optimal dispersers. *Combinatorica*, 22(1):123–145, 2002. 177

[TBR15] Gongguo Tang, Badri Narayan Bhaskar, and Benjamin Recht. Near minimax line spectral estimation. *Information Theory, IEEE Transactions on*, 61(1):499–512, 2015. 3

[Tre01] Luca Trevisan. Extractors and pseudorandom generators. *J. ACM*, 48(4):860–879, July 2001. 10

[Tro12] Joel A. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12:389–434, 2012. 18

[Tul09] Madhur Tulsiani. Csp gaps and reductions in the lasserre hierarchy. In *Proceedings of the Forty-first Annual ACM Symposium on Theory of Computing*, STOC '09, pages 303–312, New York, NY, USA, 2009. ACM. 187, 191

[TZ04] Amnon Ta-Shma and David Zuckerman. Extractor codes. *IEEE Transactions on Information Theory*, 50(12):3015–3025, 2004. 177

[Vaz86] Umesh Vazirani. Randomness, adversaries and computation. In *Ph.D. Thesis, EECS, UC Berkeley*, pages 458–463. 1986. 107

[Vaz87] U. Vazirani. Efficiency considerations in using semi-random sources. In *Proceedings of the Nineteenth Annual ACM Symposium on Theory of Computing*, STOC '87, pages 160–168, 1987. 81

[Voc03] Berthold Vocking. How asymmetry helps load balancing. *J. ACM*, 50(4):568–589, July 2003. 11, 12, 105, 106, 108, 109, 110, 122, 124

[Woe99] Philipp Woelfel. Efficient strongly universal and optimally universal hashing. In *International Symposium on Mathematical Foundations of Computer Science, MFCS, 1999*, pages 262–272, 1999. 81

[Woo14] David P Woodruff. Sketching as a tool for numerical linear algebra. *arXiv preprint arXiv:1411.4357*, 2014. 6

[WZ99] Avi Wigderson and David Zuckerman. Expanders that beat the eigenvalue bound: Explicit construction and applications. *Combinatorica*, 19(1):125–138, 1999. 177

[Ye01] Yinyu Ye. A .699-approximation algorithm for Max-Bisection. *Math. Program.*, 90(1):101–111, 2001. 13

[Zuc96a] David Zuckerman. On unapproximable versions of np-complete problems. *SIAM J. Comput.*, 25(6):1293–1304, 1996. 177

[Zuc96b] David Zuckerman. Simulating BPP using a general weak random source. *Algorithmica*, 16(4/5):367–391, 1996. 177

[Zuc07] David Zuckerman. Linear degree extractors and the inapproximability of max clique and chromatic number. *Theory of Computing*, 3(1):103–128, 2007. 8, 177, 178